ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

КНИГА 1 – МАТЕМАТИКА И МЕХАНИКА

КНИГА 2 – ПРИЛОЖНА МАТЕМАТИКА И ИНФОРМАТИКА

Том 89

ANNUAIRE

DE

L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE

LIVRE 1 – MATHEMATIQUES ET MECANIQUE

LIVRE 2 – MATHEMATIQUES APPLIQUEE ET INFORMATIQUE

Tome 89

СОФИЯ • 1998

СОФИЯ • 1998

ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Книга 1 — МАТЕМАТИКА И МЕХАНИКА

Том 89
1995

ANNUAIRE

DE

L'UNIVERSITE DE SOFIA "ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE

Livre 1 — MATHEMATIQUES ET MECANIQUE

Tome 89
1995

СОФИЯ • 1998 • SOFIA
УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
PRESSES UNIVERSITAIRES "ST. KLIMENT OHRIDSKI"

Annuaire de l' Université de Sofia "St. Kliment Ohridski"
Faculté de Mathématiques et Informatique

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 Blvd J. Bourchier, P.O. Box 48
BG-1164 Sofia, Bulgaria

Fax xx(359 2) 687 180
electronic mail:
annuaire@fmi.uni-sofia.bg

Aims and Scope. The *Annuaire* is the oldest Bulgarian journal, founded in 1904, devoted to pure and applied mathematics, mechanics and computer sciences. It is reviewed by *Zentralblatt für Mathematik*, *Mathematical Reviews* and the Russian *Referativnii Jurnal*. The *Annuaire* 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 sciences, classical, fluid and solid mechanics, and their fundamental applications.

ISSN 0205-0808

NIKOLA OBRESHKOFF (1896–1963)

On April 19 and 20, 1996, a special scientific Session commemorating the centenary of the birth of the great Bulgarian mathematician Nikola Obreshkoff (1896–1963) took place in Sofia. The Session was organized by the Faculty of Mathematics and Informatics at the "St. Kliment Ohridski" University of Sofia, the Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences

and the Union of Mathematicians in Bulgaria. The Scientific Committee of the session comprised B. Penkov, L. Davidov, T. Genchev (Chair) and D. Skordev. The Scientific Programme included six invited general lectures, devoted to the main research interests and achievements of Obreshkoff. Twenty three contributions were presented as well, most of them touching or connected to Obreshkoff's works. The full-length texts of those of the lectures, duly submitted to the Editors, are included in this volume.

The Editorial Board uses the opportunity to dedicate this volume to Nikola Obreshkoff. We think that it is the least that we can do to express, to a certain small extent, our deep appreciation for his profound influence to Bulgarian, and not only Bulgarian, mathematics. It is far beyond our ability and aims to give here a proper survey and account of Obreshkoff's numerous and deep contributions in many areas of mathematics -- a glimpse of some of them can be caught from the papers that follow. We can only add that Nikola Obreshkoff is, in our view, a spectacular example of the fact that in the realm of spirit and creativity there exist no small and no big nations.

Editorial Board

CONTENTS

Book 1

MATHEMATICS AND MECHANICS

Book 2

APPLIED MATHEMATICS AND INFORMATICS

NIKOLA OBRESHKOFF (1896–1963)

Born: March 6, 1896, Varna, Bulgaria.
University Education: 1915–1920, University of Sofia.
University Positions (University of Sofia):
Assistant professor: 1920–1921;
Associated professor: 1922–1927;
Full professor: 1928;
Head of the Chair of Algebra: 1928–1963.
Scientific Degrees:
Doctor of Mathematics of Palermo University (Italy): 1932;
Doctor of Sciences of Paris University (Sorbonne): 1933.
Academic Positions:
Member of the Bulgarian Academy of Sciences: 1945;
Director of the Institute of Mathematics at the Bulgarian Academy of Sciences: 1951–1963.
Selected Addresses:
Hamburg University, Berlin University, Geneva University, Rome University, Palermo University, Paris University (Sorbonne), Leipzig University, Dresden University.
Invited Speaker:
World Congresses of Mathematicians (Oslo 1936, Edinburgh 1958); First Congress of Slav Mathematicians (Warsaw, 1929); Congress of Balkan Mathematicians (Athens 1935); Congresses of Hungarian Mathematicians (Budapest 1950, 1960); Conference (Tagung) on Probability and Statistics (Berlin 1954); International Colloquium on Numerical Analysis (Dresden 1955).

SCIENTIFIC HERITAGE

Papers: more than 250.
Monographs:
Zeros of polynomials, Sofia 1963, Publishing House of the Bulgarian Academy of Sciences, 289 p. (Bulgarian);
Verteilung und Berechnung der Nullstellen reeller Polynome, Berlin 1963, VEB Deutscher Verlag der Wissenschaften, 298 p.;
La statistique mathématiques, Paris 1938, Herman, 66 p.;
Quelques classes de fonctions entieres limites de polynômes et de fonctions méromorphes limites de fractions rationelles, Paris 1941, Herman, 49 p.
Research Areas:
Location of Zeros, Summability of Divergent Series, Theory of Numbers, Real and Complex Analysis, Differential Equations, Numerical Analysis, Integral Geometry, Probability and Statistics, Mechanics.
Main Contributions:
— generalization of Budan-Fourier theorem and Descartes rule for complex zeros of algebraic polynomials;
— generalization of Laguerre, Poulain-Hermite and Malo theorems;
— summation of the differentiated Fourier series;
— summation of the ultraspherical series by arithmetical means;
— absolute summation by typical means;
— generalizations of Mittag-Leffler and Borel methods of summation;
— characterization of entire and meromorphic functions as limits of classes of polynomials and rational functions;
— generalization of the classical Laplace transform;
— asymptotic properties of the derivatives of functions defined on a ray of the real axis;
— solution of the problem for the exact value of the Borel constant;
— approximation of irrational numbers by continuous fractions;
— asymptotics of probability densities;
— integral geometry in the hyperbolic plane;
— generalization of Taylor formula;
— numerical methods for solution of algebraic equations.

SCIENTIFIC PROGRAMME

of the Session, dedicated to the centenary of the birth of Nikola Obreshkoff (1896–1963), Sofia, April 19–20, 1996

Invited Lectures

B. BOJANOV. On a formula of Obreshkoff.

T. GENCHEV. On the investigations of Nikola Obreshkoff connected with the regularly monotonic functions.

I. DIMOVSKY. Integral transforms in the late works of Obreshkoff.

A. OBRETENOV. The works of Obreshkoff on probability theory and mathematical statistics.

P. RUSSEV. Zeros of polynomials and entire functions in the works of Nikola Obreshkoff.

T. TONKOV. The theory of diophantine approximations and the contribution to it of Nikola Obreshkoff.

Contributed Lectures

SV. BILCHEV. Existence and uniqueness of the stationary solution of a nonlinear partial differential equations.

TS. DONCHEV, I. SLAVOV. Tikhonov's theorem for functional-differential inclusions.

M. GAVRILOV. A proof of the Gauss reciprocity law.

V. HADJIISKI. Distributions of the zeros of a sequence of the best rational approximations.

G. KARATOPRAKLIEV. On a nonlocal boundary-value problem for elliptic equations.

P. KENDEROV, V. MOORS. Fragmentability and σ-fragmentability of topological spaces.

V. KIRJAKOVA. From the integral transform of Obreshkoff to the generalized fractional calculus and the special functions.

M. MANEV. Contact conformal transformations of general type of almost contact manifolds with B-metrics. Applications.

M. MITREVA, T. STOJANOV. On certain problems of Obreshkoff.

S. MIHOVSKY. Isomorphisms and automorphisms of cross products of up-groups.

N. NACHEV. Invariants of the Silov p-subgroup of the group of normalized units of a commutative group ring with characteristics p.

N. NACHEV, T. MOLLOV. Multiplicative groups of semi-simple group algebras of Abellian p-groups over a field.

J. PANEVA-KONOVSKA. Complete systems of Bessel and inversed Bessel polynomials in spaces of holomorphic functions.

TZ. RASHKOVA. On the minimal degree of $*$-identities of antisymmetric variables in the matrix algebra of an arbitrary order with a symplectic involution $*$.

D. SKORDEV. An algorithmic approach to some problems about the representation of natural numbers as sums without repetitions.

I. SOSKOV. Constructing minimal pairs of degrees.

P. TODOROV. A simple proof of a coincidence theorem of Rubinstein and Walsh and generalizations.

A. TOMOVA. Weakened Tchebyshev's method of second order for investigating trajectories of associated dynamical systems by means of coloured fractal image's technique.

T. TONKOV. On certain properties of Klosterman's sums.

V. VIDEV. On the geometry of 4-dimensional Osserman manifolds.

N. YANEV, K. MITOV. Asymptotical laws in the theory of "recovery" connected with "some particular kinds of integral equations" considered by Obreshkoff.

S. ZLATEV, I. MAKRELOV. Iterative solution of operator equations in Banach spaces using Obreshkoff's method.

Слово, произнесено от проф. Т. Генчев при откриването на юбилейната научна сесия, посветена на стогодишнината от рождението на академик Никола Обрешков

Was du ererbt von deinen Vätern hast,
erwirb es, um es zu besitzen.

Goethe

Уважаеми колеги, скъпи гости!

Днес сме се събрали тук за да изразим нашата почит към живота и творческото дело на акад. Н. Обрешков — един от най-крупните представители на българската научна мисъл, и по този начин към съзиданието и творческото начало въобще. Задълбочените изследвания на Обрешков в различните клонове на анализа и в теорията на числата обогатиха нашата наука и му донесоха международно признание. Получил почти всички отличия, на които може да се радва истинският учен, Обрешков запази своята непосредственост и не изневери на своето призвание. Не само неговият талант, но и неговото пословично трудолюбие, подхранвано от чистата радост, която му носеше творчеството, го направиха аналитик от европейски мащаб. Той ненавиждаше помпозността и празната фразеология, а саморекламата и ламтежът за административни постове му бяха органически чужди. В поведението му ясно личеше известно дистанциране от текущия момент, толкова характерно за учените по призвание.

Научната кариера на Обрешков е впечатляваща: на двадесет и шест години е доцент, на двадесет и девет — извънреден професор, а на тридесет и две — редовен професор и ръководител на катедра. На тридесет и шест годишна възраст защитава докторат в Сорбоната, а на четиридесет и осем е академик. Обрешков е първият български математик, доказал че и тук, на наша, българска почва, даже и без специализация в чужбина, могат да се правят сериозни научни открития. Академик Наджаков ни е оставил жив спомен за силното впечатление, което Обрешков е направил на преподаватели и колеги, получавайки още като студент собствени научни резултати.

Хвърляйки поглед назад, ние с очудване констатираме, че твърде малко знаем за научната младост на Обрешков, за хората и за книгите, които са му помогнали да се формира като математик. Скромен и сдържан, той не е оставил спомени, на които да се опре евентуалният му научен биограф. Например фактът, че като млад доцент Обрешков е получил рокфелерова стипендия и през учебната 1922/1923 г. е специализирал в Берлин (вероятно при Шур), си остава почти неизвестен.

Поради някои особености на своя характер акад. Обрешков не създаде научна школа. Той обаче ни остави богато творчество, което ние, неговите духовни наследници, трябва да изучаваме и развиваме, за да не позволим на забравата да заличи неговите следи. Първата важна крачка вече е направена. От 1981 г. имаме неговия двутомник, а вчера научих, че третият том от неговите съчинения се намира в печатницата. Днешният светъл академичен празник е нова, макар и скромна стъпка в същата посока.

Като изразявам съжалението на организационния комитет, че едно внезапно, тежко заболяване попречи на проф. Ив. Чобанов да бъде между нас и да направи обзор на постиженията на Обрешков в теорията на разходящите редове, обявявам юбилейната научна сесия за открита.

София, 19 април 1996 г.

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

NIKOLA OBRESHKOFF (1896–1963)

ENCOMIUM*

BOYAN PENKOV

It is a honouring and pleasant duty to express my thanks to the organizers of this memorial meeting for having invited me to hold an introductory encomium on the occasion of the centenary of Nikola Obreshkoff, who was and continues to be a significant phenomenon in Bulgarian mathematics.

Please do not interprete my first words as an attempt to justify myself beforehand when confessing that I was confronted with obstacles, most of which pleasing but difficult to overcome. The first obstacle has been formulated by Goethe in the words of Faust:

> Ach, die Erscheinung war so riesengross,
> dass ich mich recht als Zwerg empfinden sollte.

The Bulgarian translation from 1905 of the seventeen years older colleague and friend Alexander Balabanov (another great phenomenon at the then Bulgarian horizon) of Goethes lines is as follows:

> Виденuето бе до небеса,
> а аз пред него бях играчка само.

* Invited address at the memorial meeting on the occasion of the 100th anniversary of N. Obreshkoff held at the Bulgarian Academy of Sciences on April 18, 1996. The original talk was held in Bulgarian. This is an English translation of the author.

Goethe has grasped fairly deep how difficult it is to keep the distance to somebody excelling you and yet to try to get knowing him better. So much about the first obstacle.

The second obstacle is related to time — more precisely, the time that has passed. And this kind of time is always long enough but also insufficient. Some things languish in oblivion, some others have not yet settled down to be declared history. It is now 33 years that Obreshkoff is not among us. He suddenly passed away in the late summer of 1963 and just a month later was followed by Lyubomir Tchakalov. Some colleagues called it the "black autumn" of Bulgarian mathematics. 33 years equals the age of Jesus and the span of a generation. The number of colleagues having seen Obreshkoff live can be counted today on the fingers of your hands.

And last the third obstacle. We live in a country deprived of memory. How many are our citizens who can cite the birthdates of their grandparents, how many family, municipal or institutional archives are being preserved? To commemorate people like Obreshkoff would be quite easier if there were in this country professional historians of mathematics, if such a subject was part of the curricula of the now so many math departments and was not only taught but was also an object of research. The fragmentary and pale efforts in this direction cannot fill the institutional gap. Lonely enthusiasts have repeatedly tried to change the situation (it suffices to recall the name of the late Boyan Petkantchin) but their voices faded away in the wilderness.

Obreshkoff's creative activity spans over a 40 year period, from the beginning of the twenties till the beginning of the sixties. The life of a genuine mathematician — Obreshkoff was such one *par excellence* — consists of his research results. They have been announced in about 250 publications. The average number of papers published by Obreshkoff yearly is 6 or 7, the minimum of 2 papers falls at WW2 years 1944 and 1945 and the maxima — in 1938 (10 works) and 1949 (12 works). 74 of these papers are by now collected in the first two volumes of Obreshkoff's *Collected Works* that started to appear in 1977 and stopped without any arguments in 1981, when the third volume, ready to be printed, did not reach the Publishing house of the Academy of Sciences. There is one more mystery around this edition. The first volume, out of print for a long time, was republished by the renowned editing house *Birkhäuser* in Basel together with an announcement for several further volumes. The Bulgarian mathematical community has not seen this republished volume.

Before starting the risky adventure to cast a bird's eye view over the mathematical problems that Obreshkoff has dealt with — they will be discussed in detail tomorrow at a special session at the mathematical department — let me remind you the main points of his CV.

Nikola Dimitrov Obreshkoff was born in the town of Varna on April 18th, 1896 as one of the last children in a large and bright family. The father, born in 1858, was a military officer, achieving later the rank of a colonel. The mother of ten children Kitza Obreshkova — a music lover and fluent in French, was the moral and intellectual force of the family. With the beginning of this century the family moved to the capital Sofia, where Nikola graduated in 1915 from the Second

Sofia Boys High-school. Three years earlier the 16 years old high-school student published in Vol. 8 of the *Journal of the Bulgarian Physico-mathematical Society* a paper entitled *Expressing functions of half angles through functions of whole angles*. In the fall of 1915 Nikola was admitted student in mathematics and physics at the Physico-mathematical Department of Sofia University. The First World War interrupted his studies temporarily and he served as private and later as officer in a field engineering unit. Immediately after graduating in 1920 he was appointed assistant at the Chair of Calculus. In this position he was conducting practical works with the students not only in calculus but in other subjects too, that was something common for the time, but not for nowadays. Even 25 years later most of the assistants were multipresent and worked on many different math courses, at least on two. At that time the Chair was not an organisational unit but an area for which an ordinarius (full-time professor) was responsible. Let me leave it to you to decide what kind of progress is the todays almost impossibility to ask an assistant from the Chair of Algebra to conduct practical work in calculus for, say, freshmen. Reflecting on the works of Obreshkoff, a difficult question arises: was he an algebraist or an analyst or, say, a probabilist. He was all of this together.

After two years of assistantship Obreshkoff got his 'Habilitation' in 1922 as an 'ordinary docent' (= assistant professor) with his papers on distribution of zeros of polynomials, his first love to which he remained faithful to his last gasp. One of the reviewers was Kyrill Popoff. His review reads as follows:

"Delighted by the results [of Obreshkoff] I communicated them to Prof. Dr. Issai Schur, ordinarius for higher algebra at Berlin University. Here are his impressions and his opinion on the value of the paper considered [the Habilitation schrift] expressed in a letter, which I am citing here with his kind permission:

Berlin, den 13 September 1921

Sehr gechrter Herr Kollege!

Die Arbeit des Herrn N. Obreschkoff "Über die Verteilung der Wurzeln der algebraischen Gleichungen", die Sie die Freundlichkeit hatten, mir zu überbringen, hat mich sehr interessiert. Die von Herrn Obreschkoff angegebene Erweiterung des Budan-Fourierschen Theorems auf das Komplexe Gebiet ist von bemerkenswerter Eleganz und Einfachkeit. Bedarf die Beweisführung auch noch einer erheblichen Kürzung, so zeugt die Arbeit doch von dem Scharfsinn des Verfassers und sein Resultat stellt einen wertvollen Beitrag zur Theorie der algebraischen Gleichungen dar.

Mit hochachtungsvollen Grüßen Ihr sehr ergebener

Prof. Dr. I. Schur."

And Popoff continues:

"The Habilitation schrift of Mr. Obreschkoff is a valuable contribution to the field of Higher Algebra, revealing his big talent and assuring him a leading position among the young mathematicians. It shows original thought, gift to

see by himself the fundamental issues and to achieve the solution by his own efforts. All this is demonstrated also by his paper on series, though not solving problems of the same importance as the above mentioned, it shows a formed mathematical insight and an outright individuality.

I do recommend warmly Mr. Obreshkoff for the position of assistant professor at the Chair of Higher algebra.

Dr. Kyrill Popoff
Associate Professor of
Differential and Integral Calculus."

In 1925 Obreshkoff was promoted an associate professor and in 1928 — a full-time professor and Head of the Algebra Chair. He remained at this position for 35 years.

As a young lecturer he had different courses. According to the then terminology, some 'basic' ones: Higher algebra (in two parts), Infinite series, Theory of probability, and some 'temporary': Spherical and practical astronomy, Plane analytic geometry, Differential geometry.

Obreshkoff has never been abroad for a long time as a postgraduate. His two Ph.D. degrees — from Palermo and Paris, he got in 1932 and 1933, being yet a full-time professor and author of more than thirty publications.

In order to accomplish this dry recording of facts, allow me, please, a digression. I cited above the report of Popoff, the other reviewer was Emanuel Ivanoff, the then Head of the Algebra Chair. Both reports are printed in vol. 19 (1922) of the *Annuaire de l'Université de Sofia* and thus were immediately made open to the public, together with the inaugural lecture of the newly elected professor, entitled: *Character and Problems of Algebra*. Grace to such a publicity, it is not a secret to us today who did recommend and with what arguments Obreshkoff's promotion. The responsibility which Ivanoff and Popoff assumed in 1921 we count today to their merits. In later times, especially after the forties, unfortunately things became anonimous, the reports of the reviewers being available only to a restricted circle of scholars and the original documents sinking into the archives (if not destroyed) and not made known to the general public. Try nowadays to discover who proposed whom for a certain academic position! Let us hope that the Bulgarian Academic society will realize the necessity of such a publicity, lost half a century ago.

In January 1945 Obreshkoff was elected directly as an Ordinary member of the Bulgarian Academy of Sciences and Arts, as it was called at that time. The usual path was through becoming first Corresponding member, but Obreshkoff was elected directly Ordinary member. Earlier members were Ivan Tzenoff (elected in 1929) and Lyubomir Tchakaloff (elected in 1930). I do not count here the mathematicians Vassil Vassilieff, Ivan Gyuzeleff, Emanuel Ivanoff and Georgi Kirkoff, who were members of the Bulgarian Literary Society, the ancestor of the Academy.

After the sovietisation of the Academy through the Acts of 1947 and 1949 and the foundation of some research institutes in the framework of the Academy Obreshkoff was appointed in 1950 as the first director of the recently created Mathematical Institute. The death overtook him after 13 years in this position at a

crucial moment of the institute's development. As Marshall Stone has formulated it, mathematics has turned to be not only a vocation, but a profession. Some times earlier cybernetics (as the computer sciences were called) was declared sane and removed from the 'index scientiarum prohibitorum'. Mathematical methods in the social sciences and in the humanities were accepted. The application of mathematics in industry, even the intention to go in this direction, became fashionable. Obreshkoff was not allowed to participate in this 'taw' development. I have remembered his directorship (up to 1955) by two main characteristics — at first place his absolute intolerance of low quality research, and second, his highly developed sense of responsibility concerning public affairs. And all this accompanied by an inborn allergy towards bureaucracy. It was the essence, not the form that did matter for him. He often repeated that this country is small and poor and we have to economize in everything and everywhere, and that in the field of research the then loudly proclaimed law that 'quantity develops into (of course better) quality' was false.

Dear colleagues and esteemed audience! To discuss here, even minimizing the details, the rich and diverse works of Obreshkoff is not possible and beyond my scope.

Let me try to present a very brief and general sketch. Kyrille Popoff used to say that one is working during all his life on his dissertation, literally on his 'thesis'. To some extent this applies to Obreshkoff, too. He impressed the mathematical community with his very first paper on the distribution of zeros of polynomials. His beautiful generalization of the theorem of Budan and Fourier was achieved by generalizing a lemma of Johann von Segner from the midst of the 18th century. In Segner's original proposition the factor is linear, but Obreshkoff used a quadratic one, thus generalizing Descartes' rule of signs to complex valued zeros. During the twenties and thirties of this century the distribution of values of polynomials was a busy research area and Obreshkoff was one of the prominent dramatis personae, along with Dieudonné, Faber, Féjér, Fujiwara, Kakeya, Marden, Montel, Polya, Schoenberg, Schur, Szökefalvi-Nagy, Szegö, Turan, Walsh et al. As yet mentioned, Obreshkoff did not abandon these problems until his last days. Only few months before his unexpected death two monographs were published: *Zeros of Polynomials* (in Bulgarian, Sofia) and *Verteilung und Berechnung der Nullstellen realer Polynome* (in German, Berlin). These books are the result of 40 years of active research in this field. Earlier there were published only two monographs in this area: Nr 93 of *Mémorial des sciences mathématiques* by Dieudonné (1938) and *Geometry of polynomials* by Marden (1949). The first volume of Obreshkoff's Collected Works contains 45 papers upon zeros.

In his inaugural lecture by "the basic problem of algebra" Obreshkoff meant the solution of algebraic equations. Nowadays the term "algebra" has a quite different meaning. The distribution of zeros of polynomials belongs therefore to the domain of analysis. Dieudonné's review of 1938 is called: *Théorie analytique des polynomes d'une variable.*

Obreshkoff contributed also to the distribution of zeros of entire functions, to particular meromorphic functions which are limits of special polynomials or rational functions. These results interfere with his interest in functional series and lead him

to his second great love that turned out to be very fruitful: the summation of divergent series. Unfortunately, he did not succeed to present this part of his life work in a bookform. But the bifurcations from this theory are very interesting. In his famous paper on quadrature formulae, published in the *Proceedings of the Prussian Academy of Sciences* in 1940, the approach is based on a summation formula.

During the second half of the forties Obreshkoff achieved a brilliant result in diophantine approximations and gave the answer to a problem posed by Borel as early as 1903. Obreshkoff proved that the unknown 'Borel constant' is equal to 1.

Last but not least we should not forget that Obreshkoff has interesting contributions to the probability theory (series and polynomials of Charlier connected with the Poisson distribution). They are published in the series *Actualités Scientifiques et industrielles* in 1938.

Obreshkoff must be mentioned also as the author of many and influential textbooks. In a short period the young professor published as № 93, 110 and 153 (resp. in 1930, 1932 and 1935) of the famous Bulgarian *University library* series two volumes of Higher algebra and a Collection of problems in the same field. Within 25 years the Higher algebra underwent more than five editions. But comparing the first edition (1930) with the last one (1955) you will notice the richness of the first. It contains: fundamental properties of polynomials, determinants, basic properties of algebraic equations, algebraic solution of equations, theory of numbers, theory of groups and its applications to algebraic solution, theory of Galois and finally Abel's theorem. The later editions are somehow simplified, they contain linear algebra, but some deeper topics are omitted. The second volume of this algebra textbook is in fact the first textbook on probability and statistic written by a professor of Sofia University, parallel to Oskar Anderson's (the then director of the Economical Research Institute at the University) *Einführung in die mathematische Statistik* from 1935. During the fifties the two initial volumes of the Higher algebra (which at least to me are still charming and challenging) were split among others into textbooks on probability and theory of numbers.

One can meet the name of Obreshkoff also as author of some highschool textbooks and two popular booklets (one on Euler, with co-author Yordan Duitchev, and the another under the title *What is differentiating?* with co-author Dimiter Skordev). These nice texts remind me of Herbert Robbins' joke about his coauthorship with Courant on *What is mathematics.* The version was that Courant wrote the text but put on the front page the prestigious name of young Robbins, as Hilbert did with Courant in *Methoden der mathematischen Physik.*

Tomorrow and after tomorrow during the specialised session many of you will have the possibility of following the chalk on the blackboard (the good old way to communicate mathematical ideas) to learn more on Obreshkoff's works on integral transforms and many other things. Therefore allow me to skip them here.

And now, after these words, you will be able to hear some reminiscences on the human being Obreshkoff and I shall myself not elaborate on his image that was in a moving manner unsophisticated. He had no hidden or surprising facets, but was both direct and kind. Not alien to public problems, nevertheless he was absorbed

by his internal mathematical world. I do not remember him in a bad mood, even after his physical pains became more frequent in the late fifties. He was not a lecturer for beginners but an excellent one for advanced students. This feature he had in common with Kolmogorov — they shared a creative manner of speaking and their words could be decoded only by the initiated. One more resemblance between them was that scarcely you had shared a problem you could realize they had gone through it and as Obreshkoff used to say: 'I have been thinking about this'. Indeed there were many things he had thought about.

The mathematical community of this country still owes much to Obreshkoff. We have to accomplish the edition of his complete works and we must compilate his scientific biography.

It is fine that Sofia has now an Obreshkov street, but his hospitable home at Tzar Samuel street deserves since a long time a memorial plate.

The best what future generations of Bulgarian mathematicians can do to honour the memory of Obreshkoff is to be exacting and persevering like him.

23 Oborishte str.
BG 1504 Sofia, Bulgaria
E-mail: bip@math.acad.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

ON A FORMULA OF OBRESHKOFF*

BORISLAV BOJANOV

We show that a formula given by Nikola Obreshkoff yields in a very simple way the Bernstein comparison theorem.

Keywords: devided differences, Obreshkoff formula, Bernstein comparison theorem.
1991/95 Mathematics Subject Classification: 41A03, 41A10, 41A50.

Denote by $f[x_0, \ldots, x_n]$ the *divided difference* of f at the points $x_0, \ldots, x_n$. It is well-known that if $f \in C^n[a, b]$ and $a \leq x_0 \leq \cdots \leq x_n \leq b$, then there is a point $\xi \in [x_0, x_n]$ such that

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}. \tag{1}$$

Another basic fact from calculus is the following mean value theorem: If f and g are continuously differentiable in (x, y) and $g(t) \neq 0$ for all $t \in (x, y)$, then there exists a point $\xi \in (x, y)$ such that

$$\frac{f(x) - f(y)}{g(x) - g(y)} = \frac{f'(\xi)}{g'(\xi)}. \tag{2}$$

Nikola Obreshkoff [1] has obtained a formula which extends both (1) and (2). He has exploited it to establish various inequalities for differentiable functions.

---

* Invited lecture delivered at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

The research was supported by the Bulgarian Ministry of Science under Contract No. MM–414.

Obreshkoff's formula. *Assume that f and g are from $C^{(n)}[a,b]$ and $g^{(n)}(t) > 0$ on $[a,b]$. Then for every set of points $x_0 \le \cdots \le x_n$ in $[a,b]$ there exists a point $\xi \in (x_0, x_n)$ such that*

$$\frac{f[x_0,\dots,x_n]}{g[x_0,\dots,x_n]} = \frac{f^{(n)}(\xi)}{g^{(n)}(\xi)}.$$

Proof. Set

$$A := \frac{f[x_0,\dots,x_n]}{g[x_0,\dots,x_n]}.$$

Note that $g[x_0,\dots,x_n] = g^{(n)}(t)$ for some $t \in [x_0,x_n]$ and thus $g[x_0,\dots,x_n] \neq 0$. Consider the function

$$\varphi(x) := f(x) - L_{n-1}(f;x) - A\big[g(x) - L_{n-1}(g;x)\big],$$

where $L_{n-1}(h;x)$ is the polynomial from π_{n-1} which interpolates h at $x_1,\dots,x_n$. It follows from this interpolation that $\varphi(x_i) = 0$ for $i = 1,\dots,n$. In addition, by the definition of A $\varphi(x_0) = 0$ too (because $h(x) - L_{n-1}(h;x) = h[x_1,\dots,x_n,x] \times (x-x_1)\cdots(x-x_n)$ for each function h). Thus φ has at least $n+1$ zeros. Then, by Rolle's theorem, $\varphi^{(n)}$ vanishes at a certain point $\xi \in (x_0,x_n)$, that is $\varphi^{(n)}(\xi) = f^{(n)}(\xi) - A\,g^{(n)}(\xi) = 0$ and the proof is complete.

The aim of this short note is to point out the fact that Obreshkoff's formula implies the classical Bernstein comparison theorem [2] (see also [3, Theorem 59]) concerning the best uniform polynomial approximation of a function f:

$$E_n(f) := \inf_{p\in\pi_n} \max_{x\in[a,b]} |f(x) - p(x)|.$$

Indeed, as well-known, the best approximation $E_n(f;x_0,\dots,x_{n+1})$ of f by polynomials from π_n on the finite set $x_0 < \cdots < x_{n+1}$ is related to the divided differences of f by the formula

$$E_n(f;x_0,\dots,x_{n+1}) = \left|\frac{f[x_0,\dots,x_{n+1}]}{s[x_0,\dots,x_{n+1}]}\right|,$$

where s is any function taking the values $(-1)^i$ at x_i, $i = 0,\dots,n+1$. Therefore, by Obreshkoff's formula,

$$\frac{E_n(f;x_0,\dots,x_{n+1})}{E_n(g;x_0,\dots,x_{n+1})} = \left|\frac{f^{(n+1)}(\xi)}{g^{(n+1)}(\xi)}\right|.$$

Now the following assertion is clearly true:

Assume that $f,g \in C^{(n+1)}[a,b]$ and $0 < |f^{(n+1)}(t)| \le g^{(n+1)}(t)$ for all $t \in [a,b]$. Then for each $a \le x_0 < \cdots < x_{n+1} \le b$

$$E_n(f;x_0,\dots,x_{n+1}) \le E_n(g;x_0,\dots,x_{n+1}).$$

Taking $x_0,\dots,x_{n+1}$ to be the alternating set for f, we get

$$E_n(f) = E_n(f;x_0,\dots,x_{n+1}) \le E_n(g;x_0,\dots,x_{n+1}) \le E_n(g), \tag{3}$$

which is the Bernstein comparison theorem.

Note that equality holds in (3) only if the functions f and g have a common alternating set.

REFERENCES

1. O b r e s h k o f f, N. On certain inequalities for functions of real variables. — Ann. Univ. Sofia, Fiz. Mat. Fac., **42**, 1, 1946, 213–238 (in Bulgarian).
2. B e r n s t e i n, S. N. Extremal properties of the polynomials of best approximation of a continuous function. Leningrad, 1937 (in Russian).
3. M e i n a r d u s, G. Approximation of Functions: Theory and Numerical Methods. Springer Tracts in Natural Philosophy, Vol. 13, N. Y., 1967.

Received on 30.09.1996

Department of Mathematics
University of Sofia
Blvd. James Bouchier 5
1164 Sofia, Bulgaria
E-mail: boris@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

ВЪРХУ ИЗСЛЕДВАНИЯТА НА АКАДЕМИК Н. ОБРЕШКОВ, СВЪРЗАНИ С РЕГУЛЯРНО МОНОТОННИТЕ ФУНКЦИИ[1]

ТОДОР ГЕНЧЕВ

Тодор Генчев. ОБ ИССЛЕДОВАНИЯХ АКАДЕМИКА Н. ОБРЕШКОВА, СВЯЗАННЫЕ С РЕГУЛЯРНО МОНОТОННЫМИ ФУНКЦИЯМИ

В этом статье представлен короткий обзор исследований академика Н. Обрешкова, связанные с регулярно монотонными функциями.

Todor Genchev. ON THE INVESTIGATIONS OF ACADEMICIAN N. OBRESHKOFF CONNECTED WITH REGULARLY MONOTONIC FUNCTIONS

A short survey of some investigations of the academician N. Obreshkoff connected with the regularly monotonic functions introduced by S. N. Bernstein is proposed.

В този кратък обзор ще се спра на публикациите на акад. Н. Обрешков, близки по дух с някои от класическите изследвания на С. Н. Бернщейн върху регулярно монотонните функции. Освен че в тези публикации намираме характерните за Обрешков простота и единство на методите, именно тук се съдържат и неравенствата, които привличат вниманието на младия тогава Ярослав Тагамлицки и в края на краищата го довеждат до неговата *Теорема за конусите*.

[1] Доклад, изнесен на 20 април 1996 г. на юбилейната научна сесия по случай стогодишнината от рождението на акад. Н. Обрешков.

Както ще стане ясно от самото изложение, за пряко влияние на работите на Бернщейн върху Обрешков не може да се говори. Неговите изследвания са предизвикани от естествения стремеж да си изясним връзката между някои резултати от теорията на разходящите редове, които вече е получил. Наистина на с. 105 от най-ранната му публикация [1], която може да се причисли към разглеждания цикъл, намираме следния пасаж: „При изследванията на зависимостта между условията на теоремите стигнах до някои резултати, които имат някакъв интерес". След това идва теорема 1, приведена по-долу в леко изменена редакция. Едва с течение на годините Обрешков осъзнава идейната близост на своите резултати с тези на Бернщейн и в самия край на своя жизнен и творчески път дава прости и елегантни доказателства на две от най-хубавите теореми на Бернщейн заедно с едно съществено обобщение на самото понятие за регулярно монотонна функция.

След това встъпление ще формулирам теорема 1, за която стана дума по-горе.

Теорема 1. *Нека реалните функции φ и ψ са дефинирани за $x > x_0$ и притежават непрекъснати производни до n-ти ред включително. Нека освен това е в сила неравенството*

$$|\varphi^{(n)}(x)| \leqq |\psi^{(n)}(x)|, \quad x > x_0, \tag{1}$$

и границите $\lim\limits_{x\to\infty} \varphi(x) = a$, $\lim\limits_{x\to+\infty} \psi(x) = b$ съществуват. Най-сетне нека $\psi^{(n)} \neq 0$ в целия интервал $(x_0, +\infty)$. В такъв случай е изпълнено и неравенството

$$|\varphi(x) - a| \leqq |\psi(x) - b|, \quad x > x_0. \tag{2}$$

Както отбелязва самият автор, тази теорема ни позволява да сравняваме скоростите, с които φ и ψ клонят към своите граници, когато $x \to +\infty$. За да мога да дам представа както за естеството на задачата, така и за метода на Обрешков, ще си позволя кратък коментар.

Ясно е, че за да получим (2), трябва да проинтегрираме (1) по подходящ начин. В случая $n = 1$ това се постига непосредствено. Наистина за произволни числа A и x, принадлежащи на интервала $(x_0, +\infty)$, имаме

$$|\varphi(x) - \varphi(A)| = \left|\int_A^x \varphi'(t)\,dt\right| \leqq \left|\int_A^x |\varphi'(t)|\,dt\right|$$

$$\leqq \left|\int_A^x |\psi'(t)|\,dt\right| = \left|\int_A^x \psi'(t)\,dt\right| = |\psi(x) - \psi(A)|,$$

откъдето, като оставим A да клони към $+\infty$ при фиксирано $x > x_0$, получаваме (2). От това разискване се вижда, че можем да заменим изискването $\psi' \neq 0$ с условието ψ' да не си сменя знака в интервала $(x_0, +\infty)$.

В общия случай доказателството е по-сложно, но Обрешков, прибягвайки към един от любимите си инструменти — формулата за n-тата разлика, с лекота се справя с възникналите затруднения. Ще припомня за какво става дума. Ако φ е дефинирана в интервала $(x_0, +\infty)$ за фиксирани $x > x_0$ и $h > 0$, полагаме последователно

$$\Delta_h\varphi(x) = \varphi(x+h) - \varphi(x), \quad \Delta_h^2\varphi(x) = \Delta_h\varphi(x+h) - \Delta_h\varphi(x), \quad \ldots,$$
$$\Delta_h^n\varphi(x) = \Delta_h^{n-1}\Delta_h\varphi(x)$$

и индуктивно стигаме до равенството

$$\Delta_h^n\varphi(x) = \varphi(x+nh) - \binom{n}{1}\varphi(x+(n-1)h) + \binom{n}{2}\varphi(x+(n-2)h) + \cdots + (-1)^n\varphi(x). \tag{3}$$

От друга страна, като вземем предвид (1), с помощта на класическата формула

$$\Delta_h^n\varphi(x) = \int_0^h \ldots \int_0^h \varphi^{(n)}(x+t_1+t_2+\cdots+t_n)\,dt_1dt_2\ldots dt_n$$
$$= \int_\Omega \varphi^{(n)}\Big(x+\sum_{j=1}^n t_j\Big)dt, \quad dt = dt_1dt_2\ldots dt_n, \tag{4}$$

където $\Omega \subset \mathbb{R}^n$ е n-мерният куб, дефиниран с неравенствата $0 \leqq t_j \leqq h$, $j = 1, 2, \ldots, n$, непосредствено получаваме

$$|\Delta_h^n\varphi(x)| \leqq \int_\Omega \Big|\varphi^{(n)}\Big(x+\sum_{j=1}^n t_j\Big)\Big|dt \leqq \int_\Omega \Big|\psi^{(n)}\Big(x+\sum_{j=1}^n t_j\Big)\Big|dt$$
$$= \Big|\int_\Omega \psi^{(n)}\Big(x+\sum_{j=1}^n t_j\Big)dt\Big| = |\Delta_h^n\psi(x)|,$$

защото $\psi^{(n)}$ не си сменя знака в целия интервал $(x_0, +\infty)$. По този начин Обрешков стига до решаващото съотношение

$$|\Delta_h^n\varphi(x)| \leqq |\Delta_h^n\psi(x)|, \quad x > x_0,$$

откъдето, имайки предвид (3), след граничния преход $h \to \infty$ получава неравенството

$$\Big|\varphi(x) - a\sum_{\nu=1}^n(-1)^{\nu-1}\binom{n}{\nu}\Big| \leqq \Big|\psi(x) - b\sum_{\nu=1}^n(-1)^{\nu-1}\binom{n}{\nu}\Big|, \quad x > x_0,$$

което съвпада с (2), защото очевидно $\sum_{\nu=1}^n(-1)^{\nu-1}\binom{n}{\nu} = 1$.

Втората публикация [2] от разглеждания цикъл отново е поместена в годишника на университета. Тук основен е следният резултат:

Теорема 2. *Нека f и φ са две реални функции, дефинирани за $x > a$, които притежават непрекъснати производни до n-ти ред включително и $\varphi^{(n)} \neq 0$ в целия интервал $x > a$. По-нататък, нека съществуват безкрайна редица $\{x_\nu\} \to +\infty$ и цяло число m, $0 \leqq m < n$, такива, че границите*

$$\lim_{\nu\to\infty} \frac{f(x_\nu)}{x_\nu^m} = A, \quad \lim_{\nu\to\infty} \frac{\varphi(x_\nu)}{x_\nu^m} = B \tag{5}$$

да съществуват. В такъв случай от неравенството

$$\left|f^{(n)}(x)\right| \leqq \left|\varphi^{(n)}(x)\right|, \quad x > a, \tag{6}$$

следва неравенството

$$\left|f^{(m)}(x) - m!\,A\right| \leqq \left|\varphi^{(m)}(x) - m!\,B\right|, \quad x > a. \tag{7}$$

Тази теорема е значително по-дълбока от теорема 1. В разглежданата работа Обрешков дава две доказателства: в първото си служи с едно ново представяне на n-тото нютоново частно, а във второто — с познатата формула на Монтел за същото нютоново частно, която му позволява да обхване и случая, когато f приема и комплексни стойности. За да мога да дам повече подробности, ще припомня, че ако f е функция, дефинирана в някакъв интервал (α, β), нейното n-то нютоново частно с възли $x_0 < x_1 < \ldots < x_n$, $\{x_\nu\}_0^n \subset (\alpha, \beta)$, се дефинира било чрез формулата

$$N(f, x_0, x_1, \ldots, x_n) = \sum_{\nu=0}^{n} \frac{f(x_\nu)}{P'(x_\nu)},$$

където $P(x) = \prod_{\nu=0}^{n} (x - x_\nu)$, било рекурентно чрез равенствата

$$N(f, x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0},$$

$$N(f, x_0, x_1, x_2) = \frac{N(f, x_0, x_2) - N(f, x_0, x_1)}{x_2 - x_1},$$

$$N(f, x_0, \ldots, x_m, x_{m+1}) = \frac{N(f, x_0, \ldots, x_{m-1}, x_{m+1}) - N(f, x_0, \ldots, x_{m-1}, x_m)}{x_{m+1} - x_m}.$$

Лесно се вижда [3], че ако f притежава n-та производна в (α, β), то съществува такова число $\xi \in (\alpha, \beta)$, че да имаме

$$N(f, x_0, x_1, \ldots, x_n) = \frac{f^{(n)}(\xi)}{n!}, \quad \min_\nu x_\nu < \xi < \max_\nu x_\nu. \tag{8}$$

Следователно n-тото нютоново частно на една функция с неотрицателна n-та производна е неотрицателно, както и да избираме възлите в разглеждания интервал.

Подготвяйки се за доказателството на теорема 2, Обрешков дава следното обобщение на равенство (8): Ако f и g са две функции, дефинирани и n-пъти диференцируеми в интервала (α, β), и освен това $g^{(n)}(x) \neq 0$ за $x \in (\alpha, \beta)$, то

$$\frac{N(f, x_0, x_1, \ldots, x_n)}{N(g, x_0, x_1, \ldots, x_n)} = \frac{f^{(n)}(\xi)}{g^{(n)}(\xi)}, \quad \alpha < \xi < \beta. \tag{9}$$

Доказателството на (9) не се различава от класическото и се опира на теоремата на Рол. Трябва да отбележа, че строго погледнато един междинен етап е пропуснат: най-напред би трябвало да се убедим, например с помощта на (8), че $N(g, x_0, \ldots, x_n) \neq 0$, и след това да разсъждаваме така, както прави Обрешков. Такива пропуски, които не засягат същността на доказателствата, но затрудняват четенето, намираме и на други места в богатото творчество на Обрешков. Липсата на подробно цитиране, както и на уводни бележки, които да приобщават читателя към съответната проблематика, също не спомагат за популяризирането на безспорните постижения на Обрешков.

След това отклонение да се върнем към доказателството на теорема 2. Най-напред Обрешков фиксира числата $\eta_0 < \eta_1 < \ldots < \eta_m$ произволно в интервала $x > a$ и след това избира от редицата $\{x_\nu\}$ (вж. (5)) такава подредица $\{t_i\}$, че да имаме $t_1 > \eta_k$, $0 \leqq k \leqq m$, и освен това $\lim\limits_{t_i \to \infty} \dfrac{t_{i+1}}{t_i} = \infty$. По-нататък с помощта на тъждеството

$$\left| \frac{N(f, \eta_0, \eta_1, \ldots, \eta_m, t_i, t_{i+1}, \ldots, t_{n-m+i-1})}{N(\varphi, \eta_0, \eta_1, \ldots, \eta_m, t_i, t_{i+1}, \ldots, t_{n-m+i-1})} \right| = \left| \frac{f^{(n)}(\xi)}{\varphi^{(n)}(\xi)} \right| \leqq 1$$

той стига до неравенството

$$\begin{aligned} &|N(f, \eta_0, \eta_1, \ldots, \eta_m, t_i, t_{i+1}, \ldots, t_{n-m+i-1})| \\ &\qquad \leqq |N(\varphi, \eta_0, \eta_1, \ldots, \eta_m, t_i, t_{i+1}, \ldots, t_{n-m+i-1})|, \end{aligned}$$

откъдето, като умножи с $|t_i t_{i+1} \ldots t_{i+n-m-1}|$ и извърши граничния преход $t_i \to \infty$, получава

$$|N(f, \eta_0, \eta_1, \ldots, \eta_m) - A| \leqq |N(\varphi, \eta_0, \eta_1, \ldots, \eta_m) - B|, \tag{10}$$

т. е.

$$\left| f^{(n)}(\xi) - m!\, A \right| \leqq \left| \varphi^{(n)}(\xi) - m!\, B \right|, \quad \min_k \eta_k < \xi < \max_k \eta_k. \tag{11}$$

Най-сетне, полагайки $\eta_k = x + kh$, $0 \leqq k \leqq m$, където x е произволно число от интервала $(a, +\infty)$, Обрешков оставя h да клони към нула в (11) и получава (7).

За да илюстрираме казаното, ще разгледаме простия случай $n = 2$, $m = 1$. Понеже в случая

$$N(f, x_0, x_1, x_2) = \frac{f(x_0)}{(x_1 - x_0)(x_2 - x_0)} + \frac{f(x_1)}{(x_0 - x_1)(x_2 - x_1)} + \frac{f(x_2)}{(x_0 - x_2)(x_1 - x_2)},$$

$a < x_0 < x_1 < x_2$, неравенството

$$|N(f, x_0, x_1, x_2)| \leqq |N(\varphi, x_0, x_1, x_2)|$$

взема вида

$$\left|\frac{f(x_0)}{(x_1-x_0)(x_2-x_0)}+\frac{f(x_1)}{(x_0-x_1)(x_2-x_1)}+\frac{f(x_2)}{(x_0-x_2)(x_1-x_2)}\right|$$
$$\leqq \left|\frac{\varphi(x_0)}{(x_1-x_0)(x_2-x_0)}+\frac{\varphi(x_1)}{(x_2-x_1)(x_0-x_1)}+\frac{\varphi(x_2)}{(x_0-x_2)(x_1-x_2)}\right|. \quad (12)$$

Като умножим (12) с x_2-x_0 и извършим граничния преход $x_2\to+\infty$,[1] намираме

$$\left|\frac{f(x_1)-f(x_0)}{x_1-x_0}-A\right| \leqq \left|\frac{\varphi(x_1)-\varphi(x_0)}{x_1-x_0}-B\right|,$$

откъдето, като оставим x_1 да клони към x_0, получаваме

$$|f'(x_0)-A| \leqq |\varphi'(x_0)-B|$$

и завършваме доказателството, защото x_0 е произволна точка от интервала $(a,+\infty)$.

В следващите си публикации [4–6] Обрешков опростява своя метод, като обобщава постановката на въпроса, разглеждайки и едностранни неравенства. Следващата теорема е типична.

Теорема 3 ([4]). *Нека φ и ψ са реални функции, дефинирани за $x<a$, които притежават n-ти производни, удовлетворяващи неравенството*

$$\varphi^{(n)}(x) \leqq \psi^{(n)}(x), \quad x<a. \quad (13)$$

По-нататък да предположим, че за някакво цяло m, $0\leqq m<n$, съществува редица $\{x_\nu\}\to-\infty$, за която границите

$$\lim_{x_\nu\to-\infty}\frac{\varphi(x_\nu)}{x_\nu^m}=A \quad \text{и} \quad \lim_{x_\nu\to-\infty}\frac{\psi(x_\nu)}{x_\nu^m}=B \quad (14)$$

съществуват. В такъв случай имаме

$$\varphi^{(m)}(x)-m!\,A \leqq \psi^{(m)}(x)-m!\,B, \quad x<a. \quad (15)$$

Нещо повече, ако за някакво x_0 (15) се превръща в равенство, то имаме равенство в целия интервал $x\leqq x_0$.

Ясно е, че тази теорема е по-обща от теорема 2, защото ако $\psi^{(n)}\neq 0$ за $x<a$ и е непрекъсната, можем да заменим неравенството $|\varphi^{(n)}(x)|\leqq|\psi^{(n)}(x)|$, $x<a$, с двете неравенства $\varphi^{(n)}(x)\leqq\varepsilon\psi^{(n)}(x)$ и $-\varphi^{(n)}(x)\leqq\varepsilon\psi^{(n)}(x)$, $x<a$, където ε е знакът на $\psi^{(n)}$.

Доказателството на теорема 3 се извършва по схемата, използвана за доказателството на теорема 2, но е по-просто, защото в случая е достатъчно да приложим равенство (8) към n-тото нютоново частно $N(f,\eta_0,\eta_1,\ldots,\eta_m,t_i,t_{i+1},\ldots,t_{n-m+i-1})$, където $f=\psi-\varphi$, и след граничния преход $t_i\to-\infty$ да получим

$$N(\varphi,\eta_0,\eta_1,\ldots,\eta_m)-A \leqq N(\psi,\eta_0,\eta_1,\ldots,\eta_m)-B,$$

което, както видяхме, води до (15).

[1] Предполагаме, че $x_2\to+\infty$ чрез стойности от редицата $\{x_\nu\}$, за която границите (5) съществуват.

Особено интересен е случаят $m = 0$, $A = B = 0$. Тъй като тези предположения ни осигуряват и равенствата

$$\lim_{x_\nu \to -\infty} \frac{\varphi(x_\nu)}{x_\nu^k} = 0, \quad \lim_{x_\nu \to -\infty} \frac{\psi(x_\nu)}{x_\nu^k} = 0, \quad k = 0, 1, 2, \ldots, n-1,$$

стигаме до следния забележителен резултат:

Теорема 4. *Нека функциите φ и ψ са дефинирани за $x < a$ и притежават n-ти производни, удовлетворяващи* (13). *В такъв случай, ако границите $\lim_{x_\nu \to -\infty} \varphi(x_\nu) = 0$, $\lim_{x_\nu \to -\infty} \psi(x_\nu) = 0$ съществуват за някаква редица $\{x_\nu\} \to -\infty$, то неравенствата*

$$\varphi^{(k)}(x) \leqq \psi^{(k)}(x), \quad x < a, \quad k = 0, 1, 2, \ldots, n-1 \tag{16}$$

са налице. Нещо повече, ако за някакво x_0, $x_0 < a$, и някакво ν, $0 \leqq \nu < n$, имаме $\varphi^{(\nu)}(x_0) = \psi^{(\nu)}(x_0)$, то $\varphi^{(\nu)}(x) = \psi^{(\nu)}(x)$ в целия интервал $x \leqq x_0$.

В частния случай $\psi(x) = e^x$ получаваме силно обобщение на една красива теорема на Тагамлицки [7].

Теорема 5. *Нека $n \geqq 1$ е естествено число и функцията f е n пъти диференцируема в интервала $x \leqq 0$. Ако $f(x) \leqq e^x$ за $x \leqq 0$ и освен това f удовлетворява условията $\lim_{x \to -\infty} f(x) = 0$ и $f(0) = 1$, то $f(x) = e^x$ в целия интервал $x \leqq 0$.*

Наистина според теорема 4 функцията $F(x) = e^x - f(x)$ е монотонно растяща и неотрицателна в интервала $x \leqq 0$. Понеже по условие $F(0) = 0$, то $F(x) = 0$ за $x \leqq 0$.

Ще завърша този по необходимост кратък обзор с няколко думи за изследванията на Обрешков, непосредствено свързани с теорията на регулярно монотонните функции. В своята работа [2], за която вече говорих, изхождайки от формулата на Монтел

$$N(f, x_0, x_1, \ldots, x_n) = \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} f^{(n)}(w)\, dt_n,$$

$$w = x_n t_n + x_{n-1}(t_{n-1} - t_n) + \cdots + x_0(1 - t_1), \tag{17}$$

за n-тото нютоново частно, Обрешков между другото установява следната

Теорема 6. *Нека f притежава n-та производна в интервала $x > a$ и за някаква безкрайна редица $\{x_\nu\} \to +\infty$ и някакво цяло m, $0 \leqq m < n$, границата $\lim_{x_\nu \to \infty} \frac{f(x_\nu)}{x_\nu^m} = 0$ съществува. Тогава, ако интегралът $\int_x^\infty t^{n-m-1} \left| f^{(n)}(t) \right| dt$ е сходящ за всяко $x > a$, равенството*

$$f^{(m)}(x) = m!\, A + \frac{(-1)^{n-m}}{(n-m-1)!} \int_x^\infty (t-x)^{n-m-1} f^{(n)}(t)\, dt, \quad x > a, \tag{18}$$

е удовлетворено.

Пет години по-късно Обрешков развива тази тема и публикува своята забележителна работа [8], в която между другото дава забележително просто и елегантно доказателство на една от класическите теореми на Бернщейн. Обрешков започва разискването със следната

Теорема 7. *Нека f е реална функция, дефинирана за $x > a$, която има производни до $(n+1)$-ви ред включително и удовлетворява условията*

$$(-1)^k f^{(k)}(x) \geqq 0, \quad x > a, \ k = 0, 1, 2, \ldots, n+1. \tag{19}$$

В такъв случай е в сила равенството

$$f(x) = \delta + \frac{(-1)^{n+1}}{n!} \int_x^\infty (t-x)^n f^{(n+1)}(t)\, dt, \quad x > a, \tag{20}$$

където, разбира се, $\delta = \lim\limits_{x\to\infty} f(x)$.

Именно интегралното представяне (20) е изходният пункт на Обрешков към теоремата на Бернщейн, чиято формулировка привеждам само за пълнота на изложението.

Теорема (С. Н. Бернщейн). *Нека f е реална функция, дефинирана и безбройно много пъти диференцируема в интервала $x \geqq 0$. Нека условието $(-1)^k f^{(k)}(x) \geqq 0$, $x \geqq 0$, $k = 0, 1, 2, \ldots$ е удовлетворено. В такъв случай f има вида*

$$f(x) = \int_0^\infty e^{-tx}\, d\alpha(t), \quad x > 0, \tag{21}$$

където α е дефинирана, монотонно растяща и ограничена в интервала $[0, +\infty)$.

Забележка. Функциите, удовлетворяващи условията на теоремата, се наричат *регулярно монотонни.*

Сега е моментът да скицирам доказателството на Обрешков. Без ограничение на общността можем да предположим, че $\lim\limits_{x\to\infty} f(x) = 0$, т. е. че в (20) имаме $\delta = 0$. В такъв случай след субституцията $\tau = \dfrac{n}{t}$ от (20) получаваме

$$f(x) = (-1)^{n+1} \frac{n^n}{(n-1)!} \int_0^{\frac{n}{x}} \left(1 - \frac{x\tau}{n}\right)^n \frac{1}{\tau^{n+2}} f^{(n+1)}\left(\frac{n}{\tau}\right) d\tau, \quad x > 0. \tag{22}$$

(Интегралът (22) е сходящ, защото сходимостта на (20) е установена в процеса на доказателството на теорема 7.) Остава ни да въведем монотонно растящата и ограничена функция

$$\alpha_n(\tau) = \int_0^\tau (-1)^{n+1} f^{(n+1)}\left(\frac{n}{s}\right) \frac{1}{s^{n+2}} \frac{n^n}{(n-1)!}\, ds,$$

за да представим (22) във вида

$$f(x) = \int_0^{\frac{n}{x}} \left(1 - \frac{\tau x}{n}\right)^n d\alpha_n(\tau). \tag{23}$$

Най-сетне, опирайки се на двете класически теореми на Хели, избираме сходяща подредица от $\{\alpha_n\}$ и като извършим традиционния, но деликатен граничен преход $n \to \infty$ в (23), получаваме (21) и завършваме доказателството.

Публикувано в годишника на факултета, скицираното доказателство остава незабелязано. За съжаление по същото време младият тогава съветски математик Б. Коренблюм публикува в *Успехи математических наук* [9] по същество същото доказателство, но значително по-добре редактирано и шлифовано. Препечатано от Шилов в неговия знаменит учебник [11], именно то завоюва изключителна популярност.

Тук се натъкваме на явление, което не искам да отмина с мълчание. Публикациите на Обрешков често съдържат блестящи идеи, но са твърде дълги и по правило — небрежно написани. В тях важното и второстепенното водят „мирно съвместно съществуване". Резултатът от подобна стратегия може да бъде само един — липса на популярност. Например разглежданата работа [8] е изключително богата по съдържание. Освен скицираното доказателство на теоремата на Бернщейн там намираме и елегантно доказателство на теоремата на Хаусдорф за моментите, както и скица на многомерния вариант на разглежданата теорема на Бернщейн. Наред с това обаче работата съдържа и различни варианти и отклонения, които развалят общото впечатление.

Една от последните работи на Обрешков [10] съдържа съществено обобщение на друга класическа теорема на Бернщейн, отнасяща се до регулярно монотонните функции. Ето нейната формулировка.

Теорема (С. Н. Бернщейн). *Нека реалната функция f е регулярно монотонна в крайния интервал (α, β). Тогава, каквото и да бъде числото $b \in (\alpha, \beta)$, равенството*

$$f(x) = \sum_{\nu=0}^{\infty} \frac{f^{(\nu)}(b)}{\nu!}(x-b)^\nu$$

е в сила в интервала $\alpha < x \leqq b$ и следователно ни позволява да продължим f аналитично в кръга $|z-b| < b-\alpha$.

Кратко и елегантно доказателство на тази теорема може да се намери в учебника на Тагамлицки [12].

За да обобщи този резултат на Бернщейн, Обрешков изхожда от едно сполучливо разширение на понятието регулярно монотонна функция.

Дефиниция (Н. Обрешков). Нека f е комплексна функция, дефинирана и безбройно много пъти диференцируема в крайния интервал (a, b).

Казваме, че f е *регулярно монотонна в смисъл на Обрешков*, когато са изпълнени следните изисквания:

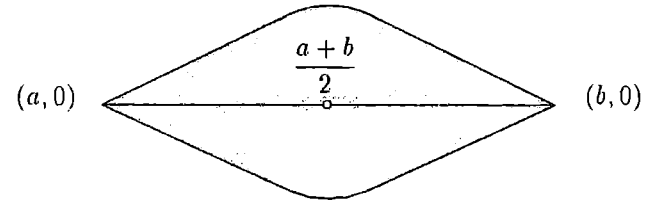
А) За всяко фиксирано $n \geqq 0$ функцията $\left|f^{(n)}(x)\right|$ е или монотонно растяща, или монотонно намаляваща в целия интервал (a, b).

Б) Когато x описва (a, b), стойностите на $f^{(n)}(x)$ лежат в някакъв ъгъл A_n с връх в началото и с големина, ненадминаваща $\pi - \delta$, където числото $\delta > 0$ не зависи от n.

Ясно е, че всяка регулярно монотонна функция е регулярно монотонна и в смисъл на Обрешков.

Сега вече мога да формулирам основния резултат на Обрешков.

Теорема 8. *Ако f е регулярно монотонна в смисъл на Обрешков в интервала (a, b), то тя се продължава аналитично поне в областта, която се получава като прекараме допирателните от точките $(a, 0)$ и $(b, 0)$ към окръжността* $\left|z - \dfrac{a+b}{2}\right| < \dfrac{b-a}{4e}$ (фиг. 1).



Фиг. 1

Доказателството на Обрешков се различава съществено от всички доказателства, дадени в реалния случай. Обрешков изхожда от едно представяне на n-тото нютоново частно, което се получава като приложим комплексния вариант на теоремата за средните стойности към интеграла в (17). (Сравнете с [13, с. 72–73].)

Доказателството на Обрешков непосредствено се обобщава за функции със стойности в крайномерни векторни пространства. За съжаление тук не мога да дам повече подробности.

ON THE INVESTIGATIONS OF NIKOLA OBRESHKOFF CONNECTED WITH THE REGULARLY MONOTONIC FUNCTIONS

(Summary)

Academician N. Obreshkoff came across this field of research studying the connections between some of his theorems about summability of a class of divergent series by typical means. His earlier result in this direction reads as follows:

Theorem 1 ([1]). *Suppose φ and ψ are real-valued functions defined for $x > x_0$ and belonging to the class $C^n(x_0, +\infty)$, where $x_0 \in \mathbb{R}$, $n \geqq 1$. Further let the limits $\lim_{x \to +\infty} \varphi(x) = a$, $\lim_{x \to +\infty} \psi(x) = b$ exist. If $\psi^{(n)} \neq 0$ in the whole interval $x > x_0$, the inequality*

$$\left|\varphi^{(n)}(x)\right| \leqq \left|\psi^{(n)}(x)\right|, \quad x > x_0, \tag{I}$$

implies the inequality

$$|\varphi(x) - a| \leqq |\psi(x) - b|, \quad x > x_0. \tag{II}$$

Obreshkoff's proof is based on the well-known integral representation of the n-th differences of φ and ψ (see (3) and (4) in the text), which leads to the decisive estimate

$$|\Delta_h^n \varphi(x)| \leqq |\Delta_h^n \psi(x)|, \quad x > x_0,\ h > 0. \tag{III}$$

Letting $h \to +\infty$ in (III), Obreshkoff completes the proof.

In the second paper [2] of this series of publications we find a deeper result.

Theorem 2 ([2]). *Let f and φ be real-valued functions in $C^n(a, +\infty)$, $a \in \mathbb{R}$, $n \geqq 1$, and let $\varphi^{(n)} \neq 0$ for $x > a$. Suppose further that there exists an infinite sequence $\{x_\nu\}_0^\infty \to +\infty$, $x_\nu > a$, and an integer m, $0 \leqq m < n$, such that the limits*

$$\lim_{\nu \to \infty} \frac{f(x_\nu)}{x_\nu^m} = A, \quad \lim_{\nu \to \infty} \frac{\varphi(x_\nu)}{x_\nu^m} = B \tag{IV}$$

exist. Then the inequality

$$\left|f^{(n)}(x)\right| \leqq \left|\varphi^{(n)}(x)\right|, \quad x > a, \tag{V}$$

implies the inequality

$$\left|f^{(m)}(x) - m!\, A\right| \leqq \left|\varphi^{(m)}(x) - m!\, B\right|, \quad x > a. \tag{VI}$$

Obreshkoff gives two proofs of this theorem. The first one uses his formula (9) (see the text) for the n-th divided differences of f and ψ, whereas his second proof is based on the Montel formula (17).

In [4–6] Obreshkoff simplifies his methods and begins considering one-sided inequalities. The following theorem is typical.

Theorem 3 ([4]). *Let φ and ψ be real-valued functions in $C^n(-\infty, a)$, $a \in \mathbb{R}$, and let the inequality*

$$\varphi^{(n)}(x) \leqq \psi^{(n)}(x), \quad x < a, \tag{VII}$$

hold. Suppose in addition that the limits

$$\lim_{x_\nu \to -\infty} \frac{\varphi(x_\nu)}{x_\nu^m} = A, \quad \lim_{x_\nu \to -\infty} \frac{\psi(x_\nu)}{x_\nu^m} = B,$$

exist for some integer m, $0 \leqq m < n$, *and for a sequence* $\{x_\nu\} \to -\infty$. *Then we have*

$$\varphi^{(m)}(x) - m!\,A \leqq \psi^{(m)}(x) - m!\,B, \quad x < a. \tag{VIII}$$

By applying Theorem 3 with $\psi(x) = e^x$, $m = 0$, $A = B = 0$, Obreshkoff obtains an interesting characterization of the exponential function.

Theorem 4. *Let* $f \in C^n(-\infty, 0]$, $n \geqq 1$, *and the inequality*

$$f(x) \leqq e^x, \quad x \leqq 0, \tag{IX}$$

is satisfied. If in addition we have $\lim_{x \to -\infty} f(x) = 0$ *and* $f(0) = 1$, *then* $f(x) = e^x$ *for* $x \leqq 0$.

After 1950 Obreshkoff's scientific interest came closer to Bernstein's subjects. In particular, in [8] we find the following

Theorem 5. *Let* f *be a real-valued function in* $C^{n+1}(a, +\infty)$ *and let*

$$(-1)^k f^{(k)}(x) \geqq 0 \quad \text{for } x > a,\ k = 0, 1, 2, \ldots, n+1. \tag{X}$$

Then the representation

$$f(x) = \delta + \frac{(-1)^{n+1}}{n!} \int_x^\infty (t-x)^n f^{(n+1)}(t)\,dt \quad (\delta = \lim_{x \to +\infty} f(x)) \tag{XI}$$

holds.

As a corollary of Theorem 5 Obreshkoff gets a simple proof of the classical Bernstein's integral representation of the regularly monotonic functions in the interval $(0, +\infty)$. Indeed, if we set $\tau = \frac{n}{t}$ in (XI) and take $\delta = 0$, we obtain (23), where $\{\alpha_n\}$ is a bounded sequence of increasing functions. By means of the well-known Helly's theorems, passing to limit in (23), Obreshkoff gets (21). Independently, at the same time a similar proof has been published by B. Korenblum in [9]. In fact, the remarkable paper [8] also contains a draft of a proof of the multidimensional version of (21), an original solution of the classical Hausdorff moment problem and of its analogue for multiple sequences as well.

In his last publication [10] Obreshkoff gives an interesting generalization of the Bernstein theorem about the analyticity of the regularly monotonic functions. In order to state the Obreshkoff's result we need a definition.

Definition. Let (a, b) be a finite interval on the real axis and let f be a complex-valued function in $C^\infty(a, b)$. We say that f is regularly monotonic in Obreshkoff's sense if it has the following properties:

a) For any $n \geqq 0$ the function $x \to \left|f^{(n)}(x)\right|$ is either increasing or decreasing in (a, b).

b) For any $n \geqq 0$ there exists an angle A_n with a vertex at the origin of the complex plane $\mathbb{C}$ and with a magnitude $|A_n| \leqq \pi - \delta$, $\delta > 0$ (δ does not depend on n). such that when x varies in (a, b), all the values of $f^{(n)}$ lie in A_n.

Now we state the last result of Obreshkoff.

Theorem 6 ([10]). *If f is regularly monotonic in (a,b) in Obreshkoff's sense, it is analytic in the domain D, $D \subset \mathbb{C}$, enclosed by two arcs of the circle $\left|z - \frac{a+b}{2}\right| < \frac{b-a}{4e}$ and four segments of the tangents to that circle passing through $(a,0)$ and $(b,0)$, respectively* (Fig. 1).

ЛИТЕРАТУРА

1. Обрешков, Н. Върху сумирането на редовете с типичните средни. — Год. на Соф. унив., Физ.-мат. фак., **41**, кн. 1, 1944–1945, 103–141.
2. Обрешков, Н. Върху някои неравенства за функциите на реални променливи. — Год. на Соф. унив., Физ.-мат. фак., **42**, кн. 1, 1945–1946, 213–238.
3. Гельфонд, А. О. Исчисление конечных разностей. Физматгиз, М., 1959.
4. Obreshkoff, N. Sur quelques inéqualités pour les differences des fonctions d'une variable réelle. — C. R. de l'Acad. Sci. Paris, **224**, 1947, 880–882. (Препечатана в „Съчинения“, т. 2, 226–227.)
5. Обрешков, Н. О некоторых неравенствах для разностей от функций действительного переменного. — ДАН СССР, **59**, № 8, 1948, 1399–1401. (Препечатана в „Съчинения“, т. 2, 287–289.)
6. Обрешков, Н. Нови неравенства за разликите на редиците и на функциите и производните им. — Известия КНК, **3**, № 1, 1947, 1–40.
7. Тагамлицки, Я. Функции, които удовлетворяват известни неравенства върху реалната ос. — Год. на Соф. унив., Физ.-мат. фак., **42**, кн. 1, 1945–1946, 239–256.
8. Обрешков, Н. Върху някои класи от функции. — Год. на Соф. унив., Физ.-мат. фак., **51**, кн. 1, 1951, 237–258. (Препечатана в „Съчинения“, т. 2, 319–334.)
9. Кореноблюм, Б. И. О двух теоремах из теории абсолютно монотонных функциях. — Успехи мат. наук, том IV, вып. 4, 1951, 172–175.
10. Обрешков, Н. Върху регулярно-монотонните функции. — Год. на Соф. унив., Физ.-мат. фак., **56**, кн. 1, 1961–1962, 27–33.
11. Шилов, Г. Е. Математический анализ, специальный курс. М., 1961.
12. Тагамлицки, Я. Диференциално смятане. IV изд., Наука и изкуство, С., 1967, 311–312.
13. Goursat, E. Cours d'Analyse Mathematique. T. II, Paris, 1911.

Received on 27.06.1996

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

ZEROES OF POLYNOMIALS AND ENTIRE FUNCTIONS IN THE WORKS OF N. OBRESHKOFF*

PETER RUSEV

In the paper some of the most remarkable Obreshkoff's results about zero distribution of algebraic polynomials and entire functions of exponential type are discussed.

Keywords: zeroes of polynomials and entire functions, Obreshkoff theorems.
Mathematics Subject Classification: 01A60, 12D10, 26C10, 30C15.

INTRODUCTION

The great Bulgarian mathematician Nikola Obreshkoff (1896–1963) left a vast scientific inheritance. About 45 of his papers contain the results of his investigations on the zero distribution of algebraic polynomials and some classes of entire functions, as well as on the numerical methods for solution of algebraic equations.

N. Obreshkoff was a world-known expert with considerable contributions to the field just mentioned. To write even a brief review on his achievements, seems to be a very hard work. That is why the author of this short survey has chosen some of the most remarkable results concerning zeroes of algebraic polynomials and entire functions of exponential type. In the first place, of course, his famous generalization of the classical Descartes rule is discusssed. Further follow his generalizations of Schur's and Malo's composition theorems obtained by means of the

* Invited lecture delivered at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

generalized Poulain – Hermite theorem. Some attention is paid to his results on zero distribution of finite Fourier transforms.

1. CLASSICAL DESCARTES RULE

1.1. The classical Descartes rule gives an upper bound for the number of the positive roots of a non-constant algebraic polynomials with real coefficients. It is remarkable that this upper bound depends only on the sign-changes of the (non-zero) coefficients of the polynomials under consideration.

Let $\lambda_0, \lambda_1, \lambda_2, \ldots$ be a finite or infinite sequence of real numbers. It is said that between λ_r and λ_s $(0 \leq r < s)$ there is a *variation* iff $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_{s-1} = 0$, and moreover $\lambda_r \lambda_s < 0$.

Let

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{1.1}$$

be a real polynomial of degree $n \geq 1$. Denote by $V = V(f)$ the number of the variations in the sequence

$$a_0,\ a_1,\ a_2,\ \ldots,\ a_n \tag{1.2}$$

and let $p = p(f)$ be the number of the positive roots of f. Then the classical Descartes rule can be formulated as follows:

The number p of the positive roots of the polynomial f is not greater than the number V of the variations in the sequence of its coefficients and in any case the difference $V - p$ is an even number, i.e.

$$p = V - 2k \tag{1.3}$$

where k is a non-negative integer.

Remark. Further, by $V = V(f)$ will be named the number of the variations of the polynomial f.

1.2. Descartes rule is formulated in the last part of his book *Discours de la methode pour bien conduir sa raison, et cherche la verité dans les sciences. Plus la dioptrique, les Meteors et la Geometrie, qui sont des essais de set methode*, Laiden, 1637, namely in *la Geometrie*.

The first proof of Descartes rule for algebraic equations with only real zeroes is due to J. A. von Segner. The auxiliar statement he has used is known now as Segner's lemma, namely:

Let $c > 0$ and $\tilde{V}$ be the number of the variations of the polynomial $(x - c)f(x)$. Then $\tilde{V} = V + 2k + 1$, where k is a non-negative integer.

Descartes rule had been formulated, proved, as well as rediscovered by many authors. Among them are J. Newton (*Universal arythmetic*, 1728), J. P. de Guadet Malv (1747), J. B. J. Fourier (1796) and F. I. Budan (1803). In the whole generality it had been proved by K. F. Gauss (1828).

Remark. The above historical data are taken from the Bulgarian translation of A. P. Jushkevitch's Comments to Descartes Geometry (Descartes, *Geometry*. Sofia, 1985, p. 199 (in Bulgarian)).

A proof, as well as numerous generalizations of Descartes rule are due to E. Laguerre (*Oeuvres*, 1, Paris, 1898).

1.3. Descartes rule is carried over equations of the kind

$$\sum_{k=0}^{n} a_k \varphi_k(x) = 0, \quad a_k \in \mathbb{R}, \ k = 0, 1, 2, \ldots, n, \tag{1.4}$$

where $\{\varphi_k(x)\}_{k=0}^{n}$ is a given system of real functions.

In the second part of G. Pólya and G. Szegö's *Aufgaben und Lehrsätze aus der Analysis*, Berlin, 1925,can be found a necessary and sufficient conditions which "ensure" the validity of Descartes rule for the equation (1.4) provided that the functions $\{\varphi_k(x)\}_{k=0}^{n}$ are sufficiently smooth.

2. BUDAN - FOURIER THEOREM

2.1. The first generalization of the classical Descartes rule is due to Budan and Fourier. Their theorem gives an upper bound for the number of the roots of a non-constant real algebraic polynomial lying in an interval of the real axis.

Let $f(x)$ be a real polynomial of degree $n \geq 1$. Then the sequence

$$f(x), f'(x), f''(x), \ldots, f^{(n)}(x), \quad x \in \mathbb{R}, \tag{BF}$$

is called Budan - Fourier (BF) sequence for the polynomial $f(x)$.

Denote by $V_x = V_x(f)$ the number of the variations in the (BF) sequence. Then the following statement is true, namely:

The number $p(a,b)$ of the roots of the polynomial f in the interval (a,b) $(a < b)$ is not greater than $V_a - V_b$ and in any case the difference $V_a - V_b - p(a,b)$ is an even number, i.e.

$$p(a,b) = V_a - V_b - 2k, \tag{2.1}$$

where k is a non-negative integer.

2.2. It is clear that Descartes rule is a particular case of Budan - Fourier theorem. Indeed, if $b > 0$ is great enough, then $V_b = 0$, i.e. $V_\infty = 0$. Moreover, since $V_0 = V$ and $p(0,\infty) = p$, the equality (1.3) is a corollary of (2.1).

3. OBRESHKOFF'S GENERALIZATION OF BUDAN - FOURIER THEOREM

3.1. Let $a < b$ and $f(x)$ be a real polynomial of degree $n \geq 1$. Denote by $M(a,b)$ the inside of the rectangle which is determined by the following conditions:

(I) It is symmetrically situated with respect to the real axis.

(II) Two of its opposite vertices are at the points a and b.

(III) The angles at these points are equal to $2\pi/(n-V_a)$ and $2\pi/V_b$, respectively.

Remark. If $V_b = 0$, i.e. when b is great enough, then $M(a,b)$ is an angular domain with a vertex at the point a.

Let further $\mu(a,b)$ be the number of the roots of the polynomial $f(x)$ in $M(a,b)$. Then the next statement is valid.

Theorem 1 (Obreshkoff's generalization of Budan – Fourier theorem [1–3]). *Let $f(a)f(b) \neq 0$, then $\mu(a,b)$ is not greater than $V_a - V_b$ and in any case the difference $V_a - V_b - \mu(a,b)$ is even, i.e.*

$$\mu(a,b) = V_a - V_b - 2s, \tag{3.1}$$

where s is a non-negative integer.

The case $a = 0$ and $b = \infty$ gives the following statement:

Theorem 2 (Obreshkoff's generalization of Descartes rule [1–3]). *Let μ be the number of the roots of the polynomial $f(x)$ having their arguments in the interval $(-\pi/(n-V), \pi/(n-V))$. Then*

$$\mu = V - 2s, \tag{3.2}$$

where s is a non-negative integer.

Remark. The classical Descartes rule is a corollary of the above statement. Indeed, if $2q$ is the number of the non-real roots of the polynomial f in the angular domain $M = M(o,\infty)$, then $\mu = p + 2q$ and (3.2) gives that $p = V - 2(q+s)$, where $q + s$ is a non-negative integer.

Another version of Theorem 2 is the next statement.

Theorem 3 (Obreshkoff [4]). *If the real polynomial f of degree $n \geq 1$ has p roots with arguments in the interval $(-\pi/(n+2-p), \pi/(n+2-p))$, then the number V of its variations is at least equal to p and moreover, the difference $V - p$ is an even number, i.e. $V = p + 2k$, where k is a non-negative integer.*

Let us mention that Theorem 1 is proved by the aid of two statements, where each of them can be regarded as analogous to Segner's lemma. Let again $f(x)$ be a real polynomial of degree $n \geq 1$ and let V be the number of its variations.

Lemma 1 (Obreshkoff [1, 3, 5]). *Let $\rho > 0$ and $0 \leq \varphi < \pi/(n+2-V)$, then the number of the variations of the polynomial $(x^2 - 2\rho\cos\varphi.x + \rho^2)f(x)$ is equal to $V + 2(k+1)$, where k is a non-negative integer.*

Lemma 2 (Obreshkoff [1, 3, 5]). *If $\rho > 0$ and $0 \leq \varphi < \pi/(V+2)$, then the number of the variations of the polynomial $(x^2 + 2\rho\cos\varphi.x + \rho^2)f(x)$ is equal to $V - 2k$, where k is a non-negative integer.*

4. SCHOENBERG'S EXTENSION OF DESCARTES RULE TO THE COMPLEX DOMAIN

A corollary of Theorem 2 is the following statement:

Let f be a real polynomial of degree $n \geq 1$ and let V be the number of its variations. Then the number ν of its roots with arguments in the interval $(-\pi/n, \pi/n)$ is not greater than V and differs from V by an even number, i.e. $\nu = V - 2k$, where k is a non-negative integer.

The first attempt to generalize the above corollary to polynomials with arbitrary complex coefficients is due to I. J. Schoenberg (*Extension of theorems of Descartes and Laguerre to the complex domain.* — Duke Math. J., **2**, 1936, 84–94). In order to formulate his result we need some definitions.

Let A be an open and convex angular domain with vertex at the origin. Define C to be its opposite angular domain, i.e. $C := \{z \in \mathbb{C} : -z \in A\}$. Both A and C form a pair of sectors, which we denote by $S = (A, C)$.

The complement of $A \bigcup C$ with respect to the complex plane is a union of two closed angular domains B and D, each of them being the opposite of the other. Let $B^* = B \setminus \{0\}$ and $D^* = D \setminus \{0\}$.

Let $F(z) = c_0 + c_1 z + c_2 z^2 + \ldots + c_n z^n$ be a non-constant polynomial with arbitrary complex coefficients. If there exists a pair of sectors $S = (A, C)$ such that all its coefficients are in $B \cup D$, then we say that S is a dividing pair of sectors for the polynomial F.

If $0 \leq r < s$ and $c_r \in B^*$, $c_s \in D^*$ or $c_r \in D^*$, $c_s \in B^*$, and moreover $c_{r+1} = c_{r+2} = \ldots = c_{s-1} = 0$, then we say that there is a *variation* between c_r and c_s. We denote the number of the variations by $V(F, S)$ in order to emphasize that it depends on the polynomial F, as well as on the dividing pair of sectors S.

Schoenberg's extension of Descartes rule is the following statement:

Let there exist a dividing pair of sectors $S(A, C)$ for the polynomial F and let $\theta \in (0, \pi)$ be the angular measure of A. Then the number of the roots of F having their arguments in the interval $(-\theta/n, \theta/n)$ is not greater than $V(F, S)$.

A refinement of the above theorem is given later by N. Obreshkoff [6].

5. VARIATION-DIMINISHING TRANSFORMATIONS

5.1. Let $A = (a_{ij})$ be a real $m \times n$-matrix. We say that the linear transorfmation $\mathbb{R}^n \to \mathbb{R}^m$ defined by the matrix A (or simply the matrix A) is *variation-diminishing* iff whatever the vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ be, then $V(x) \leq V(y)$, where $y = Ax$ and $V(x)$, resp. $V(y)$, is the number of the variations in the sequence $x_1, x_2, \ldots, x_n$, resp. $y_1, y_2, \ldots, y_m$.

In 1930 Schoenberg gave (*Über variationsvermindernde lineare Transformationen.* — Math. Zeitschr., **32**, 1930, 321–328) a sufficient condition for a real matrix to be variation-diminishing, namely:

If the matrix A is totally positive, i.e. all its minors are positive, then it is variation-diminishing.

Later T. Motzkin (*Beiträge zur Theorie der linearen Ungleichungen*, Dissertation, Basel, 1936) found necessary and sufficient conditions for a real matrix to be variation-diminishing.

A shorter proof was given by I. Schoenberg and A. Whitney (*A theorem on polygons in dimensions with application to variation-diminishing and cyclic variation-diminishing linear transformations.* — Compositio Math., **9**, 1951, 141–160).

It seems that the notion of variation-diminishing transformation, as well as Schoenberg's criterion have been inspired by Obreshkoff's proof of the generalized Budan – Fourier theorem, and in particular by that of Lemma 2. In fact Obreshkoff has proved that the matrix

$$\begin{pmatrix} -2\rho\cos\varphi & \rho^2 & 0 & 0 & 0\ldots0 & 0 \\ 0 & 1 & -2\rho\cos\varphi & \rho^2 & 0\ldots0 & 0 \\ \cdots\cdots\cdots & & & & & \\ 0 & 0 & 0 & 0 & 0\ldots1 & -2\rho\cos\varphi \end{pmatrix}$$

is variation-diminishing by establishing that all its principal minors are positive.

5.2. In Obreshkoff's paper [6] by means of Schoenberg's criterion a pure algebraic proof (i.e. without using the continuity of the polynomials considered as functions of a real variable) of the classical Budan – Fourier theorem is given. In the same paper, again by the aid of Schoenberg's criterion, the following statement is proved:

Theorem 4 (Obreshkoff [6]). *Let $a \in \mathbb{R}$, $a_k \in \mathbb{R}$, $k = 0, 1, 2, \ldots, n$, and $h \geq 0$. Then the number of the roots of the polynomial*

$$a_0+a_1(x-a)+a_2(x-a)(x-a-2h)+\cdots+a_n(x-a)(x-a-nh)^{n-1}, \quad n \geq 0,\ a_n \neq 0,$$

is less or equal to the number of the variations in the sequence $a_0, a_1, \ldots, a_n$.

The last sequence can be replaced by the sequence

$$f(a), f'(a+h), f''(a+2h), \ldots, f^{(n)}(a+nh).$$

Remark. If $a = h = 0$, then as a corollary of the above theorem one gets again the classical Descartes rule.

6. COMPOSITION THEOREMS

6.1. Let

$$A(z) = a_0 + \binom{n}{1} a_1 z + \binom{n}{2} a_2 z^2 + \ldots + a_n z^n,$$

$$B(z) = b_0 + \binom{n}{1} b_1 z + \binom{n}{2} b_2 z^2 + \ldots + b_n z^n$$

be polynomials of degree not greater than n and with arbitrary complex coefficients. Let us form the polynomial

$$C(z) = a_0 b_0 + \binom{n}{1} a_1 b_1 z + \binom{n}{2} a_2 b_2 z^2 + \ldots + a_n b_n.$$

It is of great importance to know how the distribution of the zeroes of the polynomial $C(z)$ in the complex plane depends on the distribution of the zeroes of $A(z)$ and $B(z)$.

The most popular statement answering the above question is due to G. Szegö (*Bemerkungen zu einem Satz von J. H. Grace über die Wurzeln algebraischer Gleichungen.* — Mathem. Zeitschr., **13**, 1922, 28–55), namely:

Let the zeroes of $A(z)$ be in a circular domain K and $\beta_1, \beta_2, \ldots, \beta_n$ be the zeroes of (B). Then every zero of $C(z)$ has the form $-\lambda\beta_s$, where $\lambda \in K$ and s is some of the numbers $1, 2, 3, \ldots, n$.

Remark. A circular domain in the complex plane is either the closure of the inside or the closure of the outside of a circle, or the closure of a half-plane.

The above theorem of Szegö is a corollary of a statement known as the theorem of Grace (*The zeroes of a polynomial.* — Proc. Cambridge Philos. Soc., **11**, 1902, 352–357). In fact Szegö has given to the Grace's theorem a form which is more convenient for applications.

Here are two statements which can be proved by using Szegö's theorem. The first one is due to I. Schur (*Zwei Sätze über algebraische Gleichungen mit lauter reellen Wurzeln.* — J. reine u, angew. Math., **144**, 1914, 75–88) and the second to E. Malo (*Note sur les équations algébriques dont toutes les racines sont réelles.* — J. de Math. spéciales (4), **4**, 1895, 7–10):

(I) *Let the real polynomial*

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

have only real roots and let the real polynomial

$$g(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n$$

have either only real and positive or real and negative roots. Then the polynomial

$$a_0 b_0 + 1! a_1 b_1 x + 2! a_2 b_2 x^2 + \ldots + k! a_k b_k x^k, \tag{6.1}$$

where $k = \min(m, n)$, has only real roots.

(II) *Under the same conditions and notations the polynomial*

$$a_0 b_0 + a_1 b_1 x + a_2 b_2 x^2 + \ldots + a_k b_k x^k$$

has only real roots.

6.2. The following statements generalize Schur's and Malo's theorems:

Theorem 5 (Obreshkoff [7–9]). *Let the polynomial $f(x)$ have only real zeroes and let the zeroes of the real polynomial $g(x)$ lie in the angular domain $G(m)$ defined by the inequality $|\sin\theta| \leq m^{-1/2}$ ($\theta = \arg z$). Then the polynomial* (6.1) *has only real zeroes.*

Theorem 6 (Obreshkoff [7–9]). *Let the zeroes of the both real polynomials $f(x)$ and $g(x)$ lie in the domain $G(m)$. If all the coefficients of $g(x)$ or $g(-x)$ have the same sign, then the polynomial* (6.2) *has only real zeroes.*

A classical result due to Ch. Hermite (*Questions.* — Nouv. Ann. Math., 2 sér., **5**, 1866, 432–479) and S. J. Poulain (*Théoremes géneraux sur les équations algébriques.* — Nouv. Ann. Math., 2 sér., **6**, 1867, 21–33) is the following statement:

If the polynomials $f(x)$ and $g(x)$ have only real zeroes, then so does the polynomial $g(D)f(x)$, where $D = \dfrac{d}{dx}$.

A generalization of Hermite – Poulain theorem is given by the next statement.

Theorem 7 (Obreshkoff [7–9]). *Let the polynomial $f(x)$ of degree m have only real zeroes and let the zeroes of the real polynomial $g(x)$ lie in the domain $G(m)$. Then the polynomial $g(D)f(x)$ has only real zeroes.*

The above theorem is a simple corollary of the following lemma:

Lemma 3 (Obreshkoff [7–9]). *If the polynomial $f(x)$ of degree m has only real zeroes and, moreover, $|\sin\theta| \leq m^{-1/2}$, then the polynomial*

$$f(x) - 2\rho\cos\theta . f'(x) + \rho^2 f''(x), \quad \rho > 0,$$

has only real zeroes.

Let us mention that the generalized Schur's and Malo's theorems are proved in [7–9] by means of Theorem 7.

7. ZEROES OF FINITE FOURIER TRANSFORMS

A well-known fact is that the entire functions of exponential type defined as finite Fourier transforms, namely

$$F(z) = \int_{-a}^{a} \varphi(t)\exp(izt)\,dt, \tag{7.1}$$

where $0 < a < \infty$ and $\varphi \in L_1(-a, a)$, play an important role in the mathematical analysis and its applications.

A great number of classical special functions have integral representations of the kind (7.1). A typical example is the Poisson formula

$$\sqrt{\pi}\,\Gamma(\nu + 1/2)\Big(\frac{z}{2}\Big)^{-\nu} J_\nu(z) = \int_{-1}^{1} (1 - t^2)^{\nu - 1/2}\exp(izt)\,dt,$$

where J_ν is the Bessel function of the first kind with index ν.

Particular cases of (7.1) are the entire functions

$$C(z) = \int_{0}^{a} \varphi(t)\cos zt\,dt \tag{7.2}$$

and

$$S(z) = \int_{0}^{a} \varphi(t)\sin zt\,dt. \tag{7.3}$$

Remark. It is clear that when studying the entire functions (7.1) it can be assumed $a = 1$.

A problem of considerable importance is that of the zero distribution of the entire functions (7.1), resp. (7.2) and (7.3). It has been studied by many authors and it seems that it is not exhausted till now. E. g. the problem of finding necessary

and sufficient conditions the entire functions (7.1), resp. (7.2) and (7.3), to have only finite number of non-real zeroes seems to be still open.

Remark. A more difficult problem is that of finding necessary and sufficient conditions an entire finction of the kind

$$\int_0^\infty \Phi(t) \cos zt \, dt \tag{7.4}$$

to have only finite number of non-real zeroes. This problem has been inspired by the fact that the Riemann's ζ-function has a representation of the kind (7.4).

G. Pólya was the first who studied systematically the zero distribution of the entire functions (7.1), resp. (7.2) and (7.3) (*Über die Nullstellen gewisser ganzer Functionen.* — Math. Zeitschr., **2**, 1918, 352–383). In order to formulate his main result, we introduce the class E of the real functions φ defined and R-integrable on the interval $[-1, 1]$ and having the property that the polynomials

$$P_n(\varphi; z) = \sum_{k=0} n\varphi\left(\frac{k}{n}\right) z^k$$

have their roots in the unit disk, provided that n is great enough. Pólya has proved that:

If the function φ is in the class E, then the entire functions $C(\varphi; z)$ and $S(\varphi; z)$ have only real zeroes.

Example. If φ is non-negative and not decreasing, then it is in the class E.

A rather surprising result concerning the zero distribution of the entire functions of the kind (7.2) and (7.3) has been established by N. Obreshkoff. It can be formulated as the following statement:

Theorem 8 (Obreshkoff [6]). *If the function $\varphi \in E$ and h is a real plynomial having all its roots in the half-plane $\operatorname{Re} z \leq 1/2$, then the entire functions $C(\varphi h; z)$ and $S(\varphi h; z)$ have only real zeroes.*

In fact Obreshkoff has proved that if $\varphi \in E$, then $\varphi h \in E$. He has succeeded to get this result by using the following statement:

Lemma 4 (Obreshkoff [6]). *Suppose that the (algebraic) polynomial $f(z)$ of degree $n \geq 1$ has all its roots in the unit disk. Then whatever the complex number γ with $\operatorname{Re}\gamma \geq -n/2$ be, all the roots of the polynomial $\gamma f(z) + zf'(z)$ are in the unit disk too.*

The above statement can be regarded as a "complex version" of an well-known theorem due to E. Laguerre, namely:

Let $f(x)$ be a real polynomial of degree n and γ be a real number outside of the interval $[-n, 0]$. Then the polynomial $\gamma f(x) + xf'(x)$ has as many real roots as the polynomial $f(x)$.

REFERENCES

1. O b r e s h k o f f, N. On the distribution of the roots of the algebraic equations. — Ann. de l'Univ. Sofia, **15-16**, 1918–1919, 1919–1920, 1–14, 1–11, 1–4 (in Bulgarian).
2. O b r e s h k o f f, N. On the roots of the algebraic equations. — Ann. de l'Univ. Sofia, **19**, 1922–1923, 43–76 (in Bulgarian).
3. O b r e s h k o f f, N. Über die Wurzeln von algebraischen Gleichungen. — Jahresber. Deutschen Math. Ver., **33**, 1924, 52–64.
4. O b r e s h k o f f, N. Generalization of Descartes theorem for imaginary roots. — Dokl. Acad. Nauk SSSR, **65**, 1952, 489–492 (in Russian).
5. O b r e s h k o f f, N. On the roots of the algebraic equations. — Ann. de l'Univ. Sofia, Phys.-math. Fac., Livre 1, **23**, 1927, 177–200 (in Bulgarian).
6. O b r e s h k o f f, N. On the zeros of the polynomials and some entire functions. — Ann. de l'Univ. Sofia, Phys.-math. Fac., Livre 1, **37**, 1940–1941, 1–115 (in Bulgarian).
7. O b r e s h k o f f, N. Sur une généralisation du théorème de Poulain et Hermite pour les zéros de polynomes réels. — C. R. Acad. Bulg. Sci., **11**, No 1, 1958, 5–8.
8. O b r e s h k o f f, N. On some theorems about the zeros of real polynomials. — Proc. Math. Inst. Bulg. Acad. Sci., **4**, No 2, 1960, 19–40 (in Bulgarian).
9. O b r e s h k o f f, N. Sur une généralization du théorème de Poulain et Hermite pour les zéros réels dez polynomes réels. — Acta Math. Sci. Hung., **12**, 1961, 175–184.

Received on 15.10.1996

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

THE CONTRIBUTION OF NIKOLA OBRESHKOFF TO THE THEORY OF DIOPHANTINE APPROXIMATION*

TONKO TONKOV

Toutes les Mathématiques peuvent
se déduire de la seule notion de nombre entier;
c'est là un fait aujourd'hui universellement admis.
Émile Borel

The results of Obreshkoff are compared with the similar or the same results of other mathematicians.

Keywords: diophantine approximations, continued fractions, convergent theorems.
1991/95 Mathematics Subject Classification: 11A55, 11D99.

1. THE THEORY OF DIOPHANTINE APPROXIMATION IN DEVELOPMENT

The theory of diophantine approximation, i.e. the approximation by rational numbers, begins with an investigation of Peter Gustav Lejeune Dirichlet (1805–1859). The prehistory begins with the first known approximation of an irrational number by a finite continued fraction, which is the first known writing by continued fraction. This was the Italian mathematician and engineer Rafael Bombelli (1526–1573) who presented the number $\sqrt{13}$ as equal to $3 + \dfrac{4}{6 + \frac{4}{6}}$ in his book Algebra, edited in Venezia in 1572, making an error of $\sqrt{13} - 3{,}6 < 0{,}006$.

* Invited lecture delivered at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

A half century later another Italian mathematician, Pietro Antonio Kataldi (1552–1626), introduced and studied continued fractions by using notations, close to the contemporary ones. In the book "Trattato del modo brevissimo di trovere la radice quadra delli numeri", edited in Bolognia in 1613, he wrote:

$$\sqrt{18} = 4.\&\frac{2}{8.\&\frac{2}{8.\&\frac{2}{8.}}}$$

or, briefly, $4.\&\frac{2}{8.}\&\frac{2}{8.}\&\frac{2}{8.}$.

This is a particular case of the formula

$$\sqrt{a^2+b} = a + \cfrac{b}{2a + \cfrac{b}{2a + \cfrac{b}{2a + \cdots}}}.$$

The first known application of continued fraction convergents for approximation by rational fractions with large numerators and denominators was made in 1625 by the German mathematician and philologist Daniel Schwenter (1585–1636). He used recurence relations. A more detailed study of the recurrence relations for the convergents was made by the English mathematician John Wallis (1616–1703) in his book "Arithmetica infinitorum", edited in 1656. In it he introduced the special term "fractiones continue fractae".

An important application of continued fractions was made by the Dutch mathematician, physicist and astronomer Christian Huygens (1629–1695) in connection with the planetary model of the solar system, exposed in Paris in 1680. The theoretical basis was described in his book "Descriptio automati planetarii", edited in 1698. Huygens gave the optimal ratio of the numbers of teeth of the gears, by which he modelled the revolutions of planets around the sun. He found that the convergents are the optimal rational fractions in the following meaning: If the real number α has an expansion in continued fraction and P_k/Q_k is its convergent with $Q > 1$, and if p/q is a rational fraction for which $(p,q) = 1$ and $q < Q_k$, then from $|\alpha - (p/q)| \leq |\alpha - (P_k/Q_k)|$ it follows that $q = Q_k$, and $p = P_k$. (A stronger result was given as late as 1877 by the English mathematician Henry John Smith (1826–1883)).

During the 18th century the theory of continued fractions was directed to the Analysis. Interesting results were given by Leonard Euler (1707–1783). He applied continued fractions in his monograph "Introductio in analysin infinitorum" (first edition — 1748). Euler showed that periodical continued fractions are equal to quadratic irrationalities. The reciprocal theorem was proved by Joseph Louis Lagrange (1736–1813). In a publication in 1798 Lagrange deduced the following relations:

$$\left|\alpha - \frac{P_k}{Q_k}\right| \leq \frac{1}{Q_k Q_{k+1}} < \frac{1}{Q_k^2} \quad \text{and} \quad \left|\alpha - \frac{P_k}{Q_k}\right| > \frac{1}{Q_k(Q_k + Q_{k+1})}. \qquad (1)$$

These relations express properties of continued fractions in themselves. In the second edition of his book "Essai sur la theorie des nombres" in 1808 Adrien Marie Legendre (1752–1833) proved that if $(p, q) = 1$ and

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}, \tag{2}$$

then p/q is a convergent to the continued fraction of α.

The theory of diophantine approximation begins with the study of the approximation of real numbers by rational fractions. The first result was deduced and proclaimed on April 14, 1842, by Lejeune-Dirichlet [2], who generalized a theorem about continued fractions and applied it in the theory of numbers. Dirichlet proved that if $\alpha_1, \ldots, \alpha_m$ are arbitrary real numbers and s is a positive integer, then there exist integer numbers $x_1, \ldots, x_m$, not all equal to 0, for which $|x_i| \leq s$, $i = 1, \ldots, m$, and integer number x_0, so that

$$|x_0 + \alpha_1 x_1 + \ldots \alpha_m x_m| < \frac{1}{s^m}.$$

The proof is very interesting and remarkable. In the contemporary literature the theorem of Dirichlet for the case $m = 1$ is usually formulated in the following form:

Theorem of Dirichlet. Let a and Q be real numbers and $Q > 1$. Then there exist integer numbers p and q such that

$$|\alpha q - p| < \frac{1}{Q} \quad \text{with} \quad 0 < q < Q. \tag{3}$$

Proof. Case I. Q is an integer. We consider the following $Q + 1$ numbers:

$$0,\ \{\alpha\},\ \{2\alpha\},\ \{3\alpha\},\ \ldots,\ \{(Q-1)\alpha\},\ 1, \tag{4}$$

where $\{x\}$ is the fractional part of x, i.e. $\{x\} = x - [x]$, and $[x]$ is the integer part of x (the greatest integer number not greater than x). These $Q+1$ numbers belong to the interval $[0, 1]$. We divide the interval $[0, 1]$ into the following Q subintervals:

$$\left[0, \frac{1}{Q}\right),\ \left[\frac{1}{Q}, \frac{2}{Q}\right),\ \ldots,\ \left[\frac{Q-2}{Q}, \frac{Q-1}{Q}\right),\ \left[\frac{Q-1}{Q}, 1\right]. \tag{5}$$

Obviously, there is at least one subinterval (5) which contains at least two numbers (4). Let they be $\{r\alpha\}$ and $\{s\alpha\}$ with integers r and s, $r > s$ for instance, and $0 \leq r \leq Q - 1$, $0 \leq s \leq Q - 1$. Their difference will be not greater than the length of any of the intervals (5), and this length equals to $1/Q$. So

$$\{r\alpha\} - \{s\alpha\} \leq \frac{1}{Q},$$

i.e.

$$|r\alpha - s\alpha - [r\alpha] + [s\alpha]| \leq \frac{1}{Q},$$

and denoting $r - s = q$, $[s\alpha] - [r\alpha] = p$, we have

$$|q\alpha - p| \leq \frac{1}{Q} \quad \text{and} \quad 0 < q = r - s < Q$$

as in (3).

Case II. Q is not an integer. Then instead of Q we use the number $Q' = Q + 1$ and proceed similarly to Case I.

With this the theorem is proved.

The main idea of Dirichlet, applied in this proof, can be expressed as the following principle:

If $n + 1$ things are put on n places, then there will be at least one place containing at least two things.

This is the famous principle of Dirichlet. Later, in 1907 Herman Minkowski [3] named this principle as "pigeonhole principle", thinking the places or subintervals as "pigeonholes".

The inequality of Dirichlet's theorem can be written in the following way:

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{Qq} < \frac{1}{q^2}. \tag{6}$$

These inequalities are similar to (1) and we can say that every real number can be approximated by a rational fraction p/q with exactness $1/q^2$. It is easy to deduce from (6) that if α is irrational, then there exist infinitely many rational fractions $\frac{p}{q}$ with $(p, q) = 1$ for which

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}. \tag{7}$$

This follows from the left inequality in (6) when Q tends to ∞ as α is irrational, so $\alpha - (p/q) \neq 0$. Inversely, if α is rational, the inequality (7) can be satisfied only for finitely many rational fractions p/q with $(p, q) = 1$. Indeed, let $\alpha = a/b \neq p/q$ and $(a, b) = 1$, $b > 0$, $q > 0$. Then $aq - bp \neq 0$ and

$$\left|\frac{a}{b} - \frac{p}{q}\right| = \frac{|aq - bp|}{bq} > \frac{1}{bq}.$$

If p/q are infinitely many, then there will be $q > b$ for some q and

$$\left|\frac{a}{b} - \frac{p}{q}\right| > \frac{1}{bq} > \frac{1}{q^2},$$

which contradicts (7).

Thus the theorem of Dirichlet shows different approximability of the rational and irrational numbers. This singularity was generalized two years later by Joseph Liouville (1809–1882) who proved in 1844 the remarkable theorem that if α is a real algebraic number of degree $n \geq 1$, then there exists a constant $C = C(\alpha)$ such that

$$\left|\alpha - \frac{p}{q}\right| > \frac{C}{q^n} \tag{8}$$

for all rational numbers p/q, $q > 0$, $p/q \neq \alpha$.

It is easy to find examples for α when (8) is not satisfied, such that these α are non-algebraic, transcendental numbers. The theorem of Liouville was continued by

A. Thue, C. L. Siegel and others, and completed finally by K. Roth in 1955, but here our aim is to follow directly the Dirichlet's theorem.

In 1891 Adolf Hurwitz (1859–1919) [4] proved that if α is irrational, then the inequality

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{\sqrt{5}\,q^2} \tag{9}$$

has infinitely many solutions in integers p, q with $(p, q) = 1$. This is not true if in (9) we substitute $\sqrt{5}$ by a greater number.

In 1895 K. Vahlen [5] proved that if p_{n-1}/q_{n-1} and p_n/q_n are two consecutive convergents of the real number α, expanded in a continued fraction, then at least one of them satisfies the inequality

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{2q^2}\,.$$

The theorem of Vahlen complements the assertion of Legendre about (2) that p/q can be only convergent.

In 1903 Émile Borel (1871–1956) [1] proved that if P_{n-2}/Q_{n-2}, P_{n-1}/Q_{n-1} and P_n/Q_n are three consecutive convergents to α, then at least one of them satisfies the inequality

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{\sqrt{5}\,q^2}\,.$$

The proof is achieved by *reductio ad absurdum.*

Let α be an arbitrary irrational number. Its expansion in a simple continued fraction has the form

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}} \tag{10}$$

or, briefly, $\alpha = [a_0; a_1, a_2, \ldots]$, where a_0 is an integer and a_i $(i = 1, 2, \ldots)$ are positive integers. (a_i — incomplete quotients of α. If α is rational, then $\alpha = [a_0; a_1, a_2, \ldots, a_n]$ for some integer $n \geq 0$.)

In 1918 M. Fujiwara [6] proved that if $n > 1$ and $a_{n+1} \geq 2$, then

$$\left|\alpha - \frac{P_i}{Q_i}\right| < \frac{2}{5Q_i^2}$$

for $i = n - 1$ or $i = n + 1$. (For more details about Diophantine approximation until 1936 see [7].)

2. TWO THEOREMS OF OBRESHKOFF ABOUT RATIONAL APPROXIMATION

Academician Nikola Obreshkoff (1896–1963) wrote 18 publications about diophantine approximations ([8–25]). In the first of them [8] and briefly in [12] he deduced a very important result, expressed by two theorems:

First theorem of Obreshkoff for rational approximation. *Let α be an arbitrary irrational number with expansion in simple continued fraction* (10). *Then at least one of the convergents P_{n-2}/Q_{n-2}, P_{n-1}/Q_{n-1} and P_n/Q_n to α satisfies the inequality*

$$\left|\alpha - \frac{P_i}{Q_i}\right| < \frac{1}{\sqrt{a_n^2 + 4}\, Q_i^2}. \tag{11}$$

Second theorem of Obreshkoff for rational approximation. *Let m be an arbitrary integer number, $m > 1$, and let E be the set of all irrational numbers whose incomplete quotients are $\leq m - 1$ and of their equivalent numbers. Let α be an arbitrary irrational number not belonging to E. Then for at least one of three consecutive convergents p/q to α we have*

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{\sqrt{m^2 + 4}\, q^2}. \tag{12}$$

The number $\sqrt{m^2 + 4}$ in (12) *can not be substituted by a greater number.*

The first theorem of Obreshkoff evidently is a nice generalization of the theorem of Borel. The proof is deduced by *reductio ad absurdum*.

These two theorems of Obreshkoff are reviewed in the international journals very modestly.

In *Mathematical Reviews* the great number theorist H. Davenport [26] wrote about the first theorem of Obreshkoff: "The author's first result is a simple generalization of Borel's theorem on three successive convergents to a continued fraction. Let

$$\theta = a_0 + \frac{1}{a_1+}\,\frac{1}{a_2+}\,\cdots$$

and let p_n/q_n be the general convergent to θ. Then the inequality

$$\left|\theta - \frac{p_i}{q_i}\right| < \frac{1}{q_i^2 (a_n^2 + 4)^{1/2}}$$

is satisfied for at least one of the three values $n-2$, $n-1$ and n".

In *Zentralblatt für Mathematik* another great number theorist K. Mahler [27] described the first theorem of Obreshkoff, showing the inequality (11).

In spite of the original and official publications of the theorems of Obreshkoff and their international reviews, these theorems were forgotten for years.

3. REDISCOVERING THE THEOREMS OF OBRESHKOFF

In 1955 Max Müller [28] proved several theorems and two of them punctually repeat the theorems of Obreshkoff, but his name is not cited. (In conversations with me, Obreshkoff said that he did not like the fact that his name was not cited.) The paper of Müller was reviewed in *Zentralblat fur Mathematik* (Bd. **64**, 1956, p. 44) by the very known J. W. S. Cassels, who wrote that "Theorems of Vahlen, Borel follow at once since $a_{n+1} > 1$, and theorems of Fujiwara if $a_{n+1} \geq 2$". In

Mathematical Reviews, vol. **16**, No 11, 1955, p. 1090, J. H. H. Chalk wrote that Müller "establishes several inequalities of which the following is typical. If $n \geq 1$ and the continued fraction has at least $n+2$ elements, then $\left|z - \frac{A_\nu}{B_\nu}\right| < \frac{1}{\sqrt{a_{n+1}^2 + 4}\, B_\nu^2}$ for at least one of the values $\nu = n-1, n, n+1$", but this is the first theorem of Obreshkoff. In *Реферативный журнал, Математика*, No 987, 1956, P. G. Kogoniya accurately described all the theorems of Müler. But nobody of these reviewers noted that Obreshkoff was the first. In 1959 F. E. G. Rodeja [29] proved a theorem, which was reviewed by the great specialist on continued fractions A. N. Hovanski [30] in the form: "Если $\frac{p_k}{q_k}$ $(k = 0, 1, 2, \ldots)$ — подходящие дроби цепной дроби, в которую разложено число α, $\alpha = (a_0, a_1, a_2, \ldots)$, то выполняется по меньшей мере одно из трех неравенств $\left|\alpha - \frac{p_m}{q_m}\right| \leq \frac{1}{\sqrt{4 + a_{k+1}^2}\, q_m^2}$, $m = k-1,\ k,\ k+1$. При этом число $\sqrt{4 + a_{k+1}^2}$ нельзя заменить большим даже при увеличении числа неравенств." Obreshkoff is not cited.

Evidently, Rodeja also rediscovered the theorem of Obreshkoff. But he added more about the exactness of the constant.

In 1966 F. Bagemihl and J. R. McLaughlin [31] proved the following theorem:

Let α is an arbitrary real number with expansion (10). Let s be a natural number (positive integer). If $a_{n-1} \geq s$ for some $n \geq 1$, then at least one of the three inequalities

$$\left|\alpha - \frac{p_i}{q_i}\right| < \frac{1}{\sqrt{s^2 + 4}\, q_j^2}, \quad j = n-1,\ n,\ n+1,$$

holds.

Evidently, this is the second theorem of Obreshkoff, but the authors do not cite it.

In 1982 Fuzhong Li [32] published certain results in Chinese, whose English summary in *Zentralblatt fur Mathematik* [33] shows full coincidence with the first theorem of Obreshkoff.

In 1983 Jingcheng Tong published a paper [34], in which he defined the number M_n from the equality $\left|\alpha - \frac{p}{q}\right| = \frac{1}{M_n q_n^2}$, and wrote: "In this paper we prove the following theorem which shows the conjugate property of the Borel theorem.

Theorem. For $n \geq 2$, at least one of M_i, $i = n-1, n, m+1$, exceeds $\sqrt{a_{n+1}^2 + 4}$; at least one of M_i, $i = n-1, n, m+1$, is less than $\sqrt{a_{n+1}^2 + 4}$."

Evidently, the first part of this theorem coincides with the first theorem of Obreshkoff and is not new. But its second part is really a new theorem of Tong. We shall call it the Theorem of Tong of 1983. This Tong's very interesting theorem completes the theorem of Obreshkoff.

In 1994 Tong [35] achieved in some sense the best improve of the first theorem of Obreshkoff by proving, with the above notations, that

$$M_n \leq \sqrt{(a_{n+1} + \mu_n)^2 + 4}$$

implies

$$(M_{n-1}, M_{n+1}) > \sqrt{(a_{n+1} + \mu_n)^2 + 4},$$

where

$$\mu_n = |\alpha_n - \beta_n|, \quad \alpha_n = [0, a_{n+2}, a_{n+3}, \ldots], \quad \beta_n = [0, a_n, a_{n-1}, \ldots, a_1].$$

But the name of Obreshkoff is not mentioned. Instead of this the reviewer Hans Kopetzky wrote in *Mathematical Reviews* [37] how to obtain the result of Müller as a particular case. Evidently, it was not known yet that "the result of Müller" is the first theorem of Obreshkoff.

4. ASYMMETRIC APPROXIMATION — ANOTHER WAY FOR REDISCOVERING THE OBRESHKOFF'S THEOREMS

In 1945 Beniamino Segre [38], using a geometrical method, proved the following theorem:

Let α be an arbitrary real number. Then for every real $\tau \geq 0$ there exist infinitely many rational fractions p/q such that

$$-\frac{1}{q^2\sqrt{1+4\tau}} < \alpha - \frac{p}{q} < \frac{\tau}{q^2\sqrt{1+4\tau}}. \tag{13}$$

A precision of this result of Segre was proposed by Nicolae Negoescu [39], but it turned out to be wrong, as remarked by R. A. Rankin [40]. In 1953–1954 W. J. LeVeque [39] proved the precise theorem. The author of the present paper has written more details about this history in [45].

In 1988 Jingcheng Tong [35] proved the following theorem:

Let $\tau \geq 0$ and let α be an irrational number with expansion (10), and let p_n/q_n, $n = 1, 2, \ldots$, be its convergents. Then among the three consecutive convergents p_i/q_i, $i = 2n-1$, $2n$, $2n+1$, $n \geq 1$, at least one satisfies the inequalities

$$-\frac{\tau}{q_i^2\sqrt{a_{2n+1}^2 + 4\tau}} < \alpha - \frac{p_i}{q_i} < \frac{1}{q_i^2\sqrt{a_{2n+1}^2 + 4\tau}}.$$

Evidently, putting $\tau = 1$, we receive a variant of the theorem of Obreshkoff.

5. THE FIRST CITATION OF THE FIRST THEOREM OF OBRESHKOFF IN THE FOREIGN LITERATURE

Very probably, it was H. Jager and C. Kraaikamp [44], in 1989, who first among the foreign mathematicians (relative to Bulgarians) cited the first theorem

of Obreshkoff. In his paper, Jager and Kraaikamp gave a new proof of the first theorem of Obreshkoff and of the Theorem of Tong of 1983.

However, the second theorem of Obreshkoff, which was rediscovered also by M. Müller, and by F. Bagemihl and J. R. McLaughlin, remaines forgotten (not counting the present paper and [45]).

6. ON THE CONSTANT OF BOREL

In his memoir of 1903, É. Borel [1] proved many theorems; one of them we cited above as the theorem of Borel, another one is the following:

Let a and b be given real numbers. Let M be an arbitrary positive number. Then there exist integer numbers x, y and z such that

$$|x| < M,\ |y| < M,\ |z| < M \quad \text{and} \quad |ax + by + z| < \frac{\theta\sqrt{a^2 + b^2 + 1}}{M^2},$$

where θ is a constant, not depending on a, b and M. In his History, L. E. Dickson [43, p. 96] called θ the constant of Borel, and wrote that it was not found. But in 1956, i.e. after 53 years, N. Obreshkoff [18] (also [20, 24]) proved that $\theta = 1$. We see that, unfortunately, the constant of Borel is not remarkable, and furthermore we shell speak about "constant of Borel" only historically.

7. OTHER OBRESHKOFF'S RESULTS ABOUT DIOPHANTINE APPROXIMATION

In his first paper [8] Obreshkoff improved not only the theorem of Borel, but also the classical inequality of Dirichlet, demonstrating the validity of the following theorem:

Let α be an arbitrary real number and let n be an arbitrary positive integer. Then there exist integer numbers x and y, for which $1 \leq x \leq n$ and

$$|\alpha x - y| \leq \frac{1}{n+1}.$$

The equality sign of the inequality is achieved only if $\alpha = d(n+1)$, where d is an arbitrary positive number with $(d, n+1) = 1$.

In the last paper [25] he generalized this theorem in the following way:

Let α be an integer > 0 and n be an integer $> a$. Then for every real α, for which $0 \leq a$, there exist at least two integer non-negative numbers x and y, for which $0 < x + y \leq n$ and

$$|\alpha x - y| \leq \frac{1}{\left[\dfrac{n+a}{n+1}\right] + 2}.$$

Moreover, the equality sign is achievable.

In some papers Obreshkoff generalized the inequality of Dirichlet for several variables. Especialy, in [23] he deduced as a consequence of his theorem the following theorem of Thue – Nagel:

Let a and b be integer numbers and m be an integer positive number. Then the congruence

$$ax + by \equiv 0 \pmod{m}$$

has always a solution in positive integer numbers x and y, for which $x^2 + y^2 > 0$ and $|x| \leq \sqrt{m}$, $|y| \leq \sqrt{m}$.

The generalization of Obreshkoff is the following:

Let $a_1, a_2, \ldots, a_k$ be k integer numbers and let m be a positive integer. Then the congruence

$$a_1x_1 + a_2x_2 + \ldots + a_kx_k \equiv 0 \pmod{m}$$

has a solution in integer numbers $x_1, x_2, \ldots, x_k$, not all equal to 0, satisfying the conditions

$$|x_p| \leq \sqrt[k]{m}, \quad p = 1, 2, \ldots, k.$$

When $k = 2$, we have the above cited theorem of Thue – Nagel.

In [15] Obreshkoff proved a theorem and H. Davenport wrote about it in *Mathematical Reviews* (vol. **12**, No 3, 1951, p. 163):

"The author proves the following simple but elegant variation of a well-known result on diophantine approximation. Let $\omega_1, \ldots, \omega_k$ be real numbers, and n a positive integer. Then there exist integers $x_1, \ldots, x_k$ (not all zero) and y, such that $0 \leq x_i \leq n$ and

$$|\omega_1x_1 + \ldots + \omega_kx_k + y| \leq N^{-1},$$

where $N = kn + 1$. The proof is by Dirichlet's principle."

Obreshkoff showed the conditions when the equality sign is achieved. The reviewer had a remark that the conditions "does not seem obvious to the reviewer".

In [23] Obreshkoff proved a more precise and general theorem:

Let us have the linear form

$$f = \sum_{\mu=1}^{n_1} a_{1\mu}x_\mu^{(1)} + \sum_{\mu=1}^{n_2} a_{2\mu}x_\mu^{(2)} + \ldots + \sum_{\mu=1}^{n_p} a_{p\mu}x_\mu^{(p)},$$

where $a_{1\mu}, a_{2\mu}, \ldots, a_{p\mu}$ are arbitrary real numbers and $n_1, n_2, \ldots, n_p$ are integer positive numbers. Let $m_1, m_2, \ldots, m_p$ also be integer positive numbers. Then there exist integer numbers $x_1^{(\nu)}, x_2^{(\nu)}, \ldots, x_{n_\nu}^{(\nu)}$, $n = 1, 2, \ldots, p$, not all zero but all non-negative or all non-positive, and integer y, for which

$$\left|x_\mu^{(\nu)}\right| \leq m_\nu, \quad 1 \leq \mu \leq n_\nu, \ 1 \leq \nu \leq p,$$

and

$$|f - y| \leq \frac{1}{M}, \tag{14}$$

where $M = (n_1m_1 + 1)(n_2m_2 + 1)\ldots(n_pm_p + 1)$.

The equality sign in (14) can be achieved.

1. Borel, É. Contribution à l'analyse arithmetique du continu. — J. de math. pure et appl., 9, 1903, 329–375.
2. Lejeune-Dirichlet, P. G. Verallgemeinerung eines Satzes aus der Lehre von den Kettenbruchen nebst einigen Anwendungen auf die Theorie der Zahlen. S. — B. preuss. Akad. Wii., 1842, 93–95.
3. Minkowski, H. Diophantische Approximationen. Leipzig, 1907.
4. Hurvitz, A. Ueber die angenaherte Darstellung der Irrationalzahlen durch rationale bruche. — Math. Ann., **39**, 1891, 279–284.
5. Vahlen, K. Ueber Naherungswerthe und Kettenbruche. — J. reine angew. Math., **115**, 1895, 221–233.
6. Fujiwara, M. Bemerkung zur Theorie der Approximationen der irrationalen Zahlen durch rationale Zahlen. — The Tohoku Math. J., **14**, 1918, 109–115.
7. Koksma, J. F. Diophantische approximationen. Berlin, 1936.
8. Обрешков, Н. Върху апроксимацията на ирационалните числа. — Год. на Соф. унив., Приридо-матем. фак., **45**, кн. 1, 1949, 179–201.
9. Obreshkoff, N. Sür l'approximation des nombres irrationnels. — C. R. Acad. Sci., Paris, **228**, 1949, 352–353.
10. Обрешков, Н. Апроксимация на n линейни форми с n неизвестни. — Год. на Соф. унив., Природо-матем. фак., **45**, кн. 1. 1949, 287–292.
11. Obreshkoff, N. Sür l'approximation diophantique lineaire. — Rendiconti Accad. Naz. Lincei, cl. Sci. Fis. Mat. Nat., **6** (8), 1949, 283–285.
12. Обрешков, Н. О приближении иррациональных чисел рациональными дробьями. — Доклады БАН, **3**, 1, 1950, 1–4.
13. Обрешков, Н. Върху диофантовите апроксимации на линейните форми при положителни стойности на променливите. — Год. на Соф. унив., Природо-матем. фак., **46**, кн. 1, 1950, 343–356.
14. Obreshkoff, N. Sür l'approximation diophantique lineaire. — C. R. Acad. Bulg. Sci., **3**, No 2–3, 1950, 1–4.
15. Обрешков, Н. О диофантовых приближениях линейных форм для положительных значений переменных. — Доклади АН СССР, **73**, No 1, 1950, 21–24.
16. Obreshkoff, N. Sür l'approximation diophantique lineaire pour des valeurs positifs des variables. — C. R. Acad. Bulg. Sci., **4**, No 1, 1951, 1–4.
17. Обрешков, Н. Върху апроксимацията на линейните форми. — Известия Мат. инст. БАН, **1**, кн. 2, 1954, 35–46.
18. Obreshkoff, N. Sür une question de l'approximation diophantique des formes lineaire. — C. R. Acad. Bulg. Sci., **9**, No 4, 1956, 1–4.
19. Обрешков, Н. Две теореми за апроксимацията на линейните форми. — Известия Мат. инст. БАН, **2**, кн. 1, 1956, 35–43.
20. Обрешков, Н. Върху някои точни неравенства за диофантовите приближения на линейните форми. — Известия Мат. инст. БАН, **2**, 1957, 19–44.
21. Obreshkoff, N. Sür l'approximation diophantienne des nombres reels. — C. R. Acad. Sci. Paris, **246**, 1958, 31–32.
22. Obreshkoff, N. Sür l'approximation diophantienne des formes lineaires. — C. R. Acad. Sci. Paris, **246**, 1958, 204–205.
23. Obreshkoff, N. Sür l'approximation diophantienne des formes lineaires. — Arkiv for Matematik, **3**, 1958, 537–542.
24. Обрешков, Н. Об одном из вопросов диофантовых приближений линейных форм. В: Труды третьего всесоюзного математического съезда. Москва, июнь–июль 1956, т. IV, М., 1959, 133–139.
25. Обрешков, Н. Върху диофантовите приближения на линейни форми. — Известия Мат. инст. БАН, **3**, кн. 2, 1959, 3–18.
26. Davenport, H. — Mathematical Reviews, **12**, No 8, 1951, p. 595.
27. Mahler, K. — Zentralblatt fur Mathematik, **43**, 1952, p. 33.

28. Müller, M. Über die Approximation reeller Zahlen durch die Naherungsbruche ihres regelmassigen Kettenbruches. — Archiv der Mathematik, **6**, 1955, 253–258.
29. Rodeja, F. E. G. Un teorema de fracciones continuas. — Rev. mat. hisp.-amer., **19**, No 5–6, 1959, 231–234.
30. Хованский, А. Н. — Реферативный журнал, Математика, **2**, А 60, 1961.
31. Bagemihl, F. J. R. McLaughlin. Generalization of some classical theorems concerning triples of consecutive convergents to simple continued fractions. — J. Reine angew. Math., **221**, 1966, 146–149.
32. Li, F. A generalization of the Hurwitz theorem. — J. Northeast Norm. Univ., Nat. Sci. Ed., **1**, 1982, 39–46.
33. — Zentralblatt fur Mathematik, **565**, 1986, No 10028, p. 42.
34. Tong, J. The conjugate property of the Borel Theorem on Diophantine Approximation. — Math. Zeitschrift, **184**, 1983, 151–153.
35. Tong, J. Segre's theorem on asymmetric Diophantine approximation. — J. of Number Theory, **28**, 1988, 116–118.
36. Tong, J. The best approximation function to irrational numbers. — J. Number Th., **49**, 1994, 89–94.
37. — Mathematical Reviews, 96b: 11101, 1996.
38. Segre, B. Lattice points in infinite domains and asymmetric Diophantine approximations. — Duke Math. J., **12**, 1945, 337–365.
39. Negoescu, N. Nota asupra unei teoreme de aproximatiuni asimetrice. — Acad. Rep. pop. Romane. Bul. Sti., A. Mat., Fiz., **1**, No 2, 1949, 1–3.
40. Rankin, R. A. — Mathematical Reviews, **13**, No 7, 1952, p. 630.
41. LeVeque, W. J. On asymmetric approximations. — The Michigan Math. J., **2**, 1953–1954, 1–6.
42. Tong, J. Segre's theorem on asymmetric Diophantine approximation. — J. of Number Theory, **28**, 1988, 116–118.
43. Dickson, L. E. History of the theory of numbers. Vol. 2, N. Y., 1920.
44. Jager, H., C. Kraaikamp. On the approximation by continued fractions. — Indag. Math., **51**, 1989, 289–307.
45. Тонков, Т. Диофантовите апроксимации в творчеството на Н. Обрешков. В: Български математици, С., 1987, 132–135.

Received on 15.07.1996

University of Mining and Geology
Department of Mathematics
1100 Sofia, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

HYPERBOLIC AND EUCLIDEAN DISTANCE FUNCTIONS*

WALTER BENZ

In memory of
Nikola Obreshkoff (1896-1963),
the great Bulgarian mathematician

This is a functional equations approach to the non-negative functions $h(x,y)$ and $e(x,y)$ as defined in formulas (1) and (2). Moreover, all distance functions of $\mathbb{R}^n$ are characterized, which are invariant under linear and orthogonal mappings (see Theorem 1), and, especially, all functions of this type are determined, which satisfy in addition (D_2) (see Theorem 2). Here (D_2) asks for the invariance under euclidean or hyperbolic translations of the x_1-axis. Finally, additivity on the x_1-axis is considered, leading to the distance functions h and e up to non-negative factors (see Theorem 3).

Keywords: hyperbolic distance, invariance of distance functions under special motions, additivity on a line. theorems.
1991 Mathematics Subject Classification: 39B40, 51M10, 51K05.

1. Let $n > 1$ be an integer and let $\mathbb{R}_{\geq 0}$ be the set of all non-negative real numbers. A function

$$d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$$

is then called a *distance function* of $\mathbb{R}_{\geq 0}$. Especially, we are interested in the *hyperbolic distance function* $h(x,y)$ satisfying

$$\cosh h(x,y) = \sqrt{1+x^2}\sqrt{1+y^2} - xy, \tag{1}$$

* Lecture accepted for the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

and, moreover, in the *euclidean distance function* $e(x,y)$ defined by

$$e(x,y) = \sqrt{(x-y)^2}. \tag{2}$$

In formulas (1) and (2)

$$uv = u_1v_1 + u_2v_2 + \cdots + u_nv_n$$

denotes the usual scalar product of the elements

$$u = (u_1, \ldots, u_n) \quad \text{and} \quad v = (v_1, \ldots, v_n)$$

of $\mathbb{R}^n$.

We will say that the distance function d of $\mathbb{R}$ is of type (D_1) if, and only if, it satisfies

(D_1) $d(x,y) = d(\varphi(x), \varphi(y))$ *for all* $x, y \in \mathbb{R}^n$ *and all linear and orthogonal mappings* φ *of* $\mathbb{R}^n$.

Obviously, distance functions h and e are of type (D_1).

2. It is posible to determine all distance functions d of $\mathbb{R}^n$ which are of type (D_1). We would like to prove the following

Theorem 1. *Define*

$$K := \left\{(\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3 \mid \xi_1, \xi_2 \in \mathbb{R}_{\geq 0} \text{ and } \xi_3^2 \leq \xi_1\xi_2\right\}.$$

Suppose that $f : K \to \mathbb{R}_{\geq 0}$ *is chosen arbitrarily. Then*

$$d(x,y) = f(x^2, y^2, xy) \tag{3}$$

is a distance function of $\mathbb{R}^n$ *of type* (D_1). *If, vice versa,* d *is a distance function of* $\mathbb{R}^n$ *of type* (D_1), *there exists* $f : K \to \mathbb{R}_{\geq 0}$ *such that* (3) *holds true for all* $x, y \in \mathbb{R}^n$.

Proof. Since $x^2 = [\varphi(x)]^2$ and $xy = \varphi(x)\varphi(y)$ for all $x, y \in \mathbb{R}^n$ and for every linear and orthogonal mapping φ of $\mathbb{R}^n$ into itself, we get

$$d(x,y) = d(\varphi(x), \varphi(y)).$$

d is hence of type (D_1).

Assume now that d is a distance function of $\mathbb{R}^n$. Suppose that

$$(\xi_1, \xi_2, \xi_3)$$

is an element of K and define

$$e_1 = (1, 0, \ldots, 0) \quad \text{and} \quad e_2 = (0, 1, 0, \ldots, 0)$$

as elements of $\mathbb{R}^n$. Put

$$x_0 = 0 \quad \text{and} \quad y_0 = e_1\sqrt{\xi_2}$$

in the case $\xi_1 = 0$. Observe here $\xi_3 = 0$, in view of $\xi_3^2 \leq \xi_1\xi_2$. Define now

$$f(\xi_1, \xi_2, \xi_3) = d(x_0, y_0).$$

Put $x_0 = e_1\sqrt{\xi_1}$ and

$$y_0 = \frac{e_1\xi_3 + e_2\sqrt{\xi_1\xi_2 - \xi_3^2}}{\sqrt{\xi_1}}$$

in the case $\xi_1 > 0$. Again define

$$f(\xi_1, \xi_2, \xi_3) = d(x_0, y_0).$$

Two things must now be proved. First of all we have to show that the function f is well-established. But since (ξ_1, ξ_2, ξ_3) is in K, there are only these two cases $\xi_1 = 0$ or $\xi_1 > 0$, and in both cases the value under f is uniquely determined. The second thing we have to prove, is that

$$d(x, y) = f(x^2, y^2, xy)$$

holds true for all $x, y \in \mathbb{R}^n$. Let x, y be elements of $\mathbb{R}^n$ and put

$$x^2 =: \xi_1, \quad y^2 =: \xi_2, \quad xy =: \xi_3.$$

Because of the Cauchy–Schwarz inequality, (ξ_1, ξ_2, ξ_3) must be an element of K. If we are able to prove that, there exists a linear and orthogonal mapping

$$\varphi : \mathbb{R}^n \to \mathbb{R}^n$$

satisfying

$$\varphi(x_0) = x \quad \text{and} \quad \varphi(y_0) = y,$$

where x_0, y_0 are the already defined elements with respect to ξ_i, then

$$d(x, y) = d(x_0, y_0) = f(\xi_1, \xi_2, \xi_3) = f(x^2, y^2, xy)$$

holds true and (3) is established. We now make use of the following simple statement: let a_1, a_2, a_3, b_1, b_2, b_3 be points of $\mathbb{R}^n$. Then there exists an orthogonal mapping ψ of $\mathbb{R}^n$ with

$$\psi(a_i) = b_i \quad \text{for all} \quad i \in \{1, 2, 3\}$$

if, and only if,

$$(a_i - a_j)^2 = (b_i - b_j)^2 \tag{4}$$

is satisfied for all $i, j \in \{1, 2, 3\}$ with $i < j$.

In order to apply this statement, we put

$$a_1 = 0 = b_1$$

and $a_2 = x_0$, $a_3 = y_0$, $b_2 = x$, $b_3 = y$. Since the assumptions (4), namely

$$x_0^2 = \xi_1 = x^2, \quad y_0^2 = \xi_2 = y^2$$

and $(x_0 - y_0)^2 = \xi_1 - 2\xi_3 + \xi_2 = (x - y)^2$ are satisfied, ψ exists; which is in addition linear in view of

$$\psi(0) = \psi(a_1) = b_1 = 0. \qquad \blacksquare$$

In the case of the hyperbolic distance function we apply the branch $\arg \geq 0$ of the inverse function of cosh and we have

$$f(x^2, y^2, xy) = \arg\left(\sqrt{1 + x^2}\sqrt{1 + y^2} - xy\right).$$

In the case of $e(x, y)$ we get

$$f(x^2, y^2, xy) = \sqrt{x^2 + y^2 - 2xy}.$$

3. We would like to prove the following statement. If $z \neq 0$ is an element of $\mathbb{R}^n$, then there exists a bijection γ of $\mathbb{R}^n$ with $\gamma(0) = z$ and

$$h(x, y) = h(\gamma(x), \gamma(y))$$

for all $x, y \in \mathbb{R}^n$.

There definitely exists a linear and orthogonal mapping φ with $\varphi(z) = e_1\sqrt{z^2}$. Take now $t \geq 0$ satisfying

$$\cosh t = \sqrt{1 + z^2}.$$

Then $\tau(x) := \left(x_1 \cosh t + \sqrt{1 + x^2} \sinh t, x_2, \ldots, x_n\right)$ must be a bijection of $\mathbb{R}^n$, transforming 0 into

$$(\sinh t, 0, \ldots, 0) = e_1\sqrt{z^2}.$$

Now put $\gamma = \varphi^{-1}\tau$ and observe that

$$h(x, y) = h(\tau(x), \tau(y))$$

holds true for all $x, y \in \mathbb{R}^n$.

Remark. For more information about the mapping τ see the book [5] of the author.

It it well-known that $\mathbb{R}^n$ is a metric space with respect to the distance function $e(x, y)$. We would like to show the following

Proposition. *$\mathbb{R}^n$ is a metric space with respect to the distance function $h(x, y)$.*

Proof. Suppose that x, y are elements of $\mathbb{R}^n$. The inequality of Cauchy–Schwarz

$$(xy)^2 \leq x^2y^2$$

then implies $(xy)^2 \leq x^2y^2 + (x - y)^2$, i.e.

$$(xy)^2 + 2xy + 1 \leq (1 + x^2)(1 + y^2)$$

and hence $xy + 1 \leq |xy + 1| \leq \sqrt{1 + x^2}\sqrt{1 + y^2}$. We thus have

$$\sqrt{1 + x^2}\sqrt{1 + y^2} - xy \geq 1,$$

so that (1) determines $h(xy) \geq 0$ uniquely. In view of (1), obviously,

$$h(x, y) = h(y, x)$$

holds true for all $x, y \in \mathbb{R}^n$. Observe, moreover, $h(x, x) = 0$ for all $x \in \mathbb{R}^n$. Suppose now that $h(x, y) = 0$. Then (1) implies

$$(xy)^2 = (x - y)^2 + x^2y^2.$$

If x were $\neq y$, we would have the contradiction

$$(xy)^2 \leq x^2y^2 < (x - y)^2 + x^2y^2.$$

In order to prove the triangle inequality

$$h(x, z) \leq h(x, y) + h(y, z), \tag{5}$$

take a bijection γ of $\mathbb{R}^n$ satisfying $\gamma(0) = y$ and

$$h(p, q) = h(\gamma(p), \gamma(q)) \tag{6}$$

for all $p, q \in \mathbb{R}^n$. Put $a = \gamma^{-1}(x)$ and $b = \gamma^{-1}(z)$. Then we shall prove

$$h(a,b) \leq h(a,0) + h(0,b), \tag{7}$$

which leads to (5) by applying (6). Now observe

$$-ab \leq |ab| \leq \sqrt{a^2}\sqrt{b^2},$$

i.e. $\sqrt{1+a^2}\sqrt{1+b^2} - ab \leq \sqrt{1+a^2}\sqrt{1+b^2} + \sqrt{a^2}\sqrt{b^2}$. Hence

$$\cosh h(a,b) \leq \cosh h(a,0).\cosh h(0,b) + \sinh h(a,0).\sinh h(0,b)$$

by observing

$$0 \leq \sinh h(a,0) = \sqrt{\cosh^2 h(a,0) - 1} = a^2$$

and $0 \leq \sinh h(0,b) = b^2$. Thus

$$\cosh h(a,b) \leq \cosh(h(a,0) + h(0,b)).$$

This implies (7) since $\cosh t_1 \leq \cosh t_2$ leads to $t_1 \leq t_2$ for non-negative real numbers t_1, t_2.

Remark. Observe that $\mathbb{R}^n$ is also a metric space under the rather strange distance function

$$d(x,y) := h(x,y) + e(x,y)$$

(for all $x, y \in \mathbb{R}^n$) which is of type (D_1) as well.

4. We shall call a distance function $d(x,y)$ an *euclidean* (or a *hyperbolic*) distance function if it admits all euclidean (or all hyperbolic) motions.

Define for a distance function d the property (D_2), as follows:

(D_2) *$d(x,y) = d(\tau(x), \tau(y))$ for all $x, y \in \mathbb{R}^n$ and all euclidean (or hyperbolic) translations of the x_1-axis.*

The euclidean translations of the x_1-axis are the mappings

$$(x_1, \ldots, x_n) \to (x_1 + t, x_2, \ldots, x_n)$$

for $t \in \mathbb{R}$; the hyperbolic translations of the same axis are the already defined mappings

$$x \to \left(x_1 \cosh t + \sqrt{1+x^2} \sinh t, x_2, \ldots, x_n\right). \tag{8}$$

Theorem 2. *Let g be a function from $\mathbb{R}_{\geq 0}$ into $\mathbb{R}_{\geq 0}$. Then*

$$d(x,y) = g(e(x,y))$$

is an euclidean distance function, and

$$d(x,y) = g(h(x,y))$$

is a hyperbolic distance function. There are no other distance functions satisfying (D_1) and (D_2).

Proof. a) Let us assume that d satisfies (D_1) and (D_2) with respect to euclidean translations. Then d admits all congruent mappings of $\mathbb{R}^n$, in view of (D_1) and (D_2). Hence

$$d(x,y) = d(x+(-y), y+(-y)) = d(x-y,0)$$

and thus $d(x,y) = f((x-y)^2,0,0)$ because of Theorem 1. Define

$$g(\xi) := f(\xi^2,0,0)$$

for all real $\xi \geq 0$. Hence

$$d(x,y) = g\left(\sqrt{(x-y)^2}\right) = g(e(x,y)).$$

b) Suppose that d is a distance function satisfying (D_1) and (D_2) with respect to hyperbolic translations. From

$$(x_1,\ldots,x_n) \in \mathbb{R}^n$$

we go over to Weierstrass co-ordinates

$$\left(x_1,\ldots,x_n,\sqrt{1+x^2}\right).$$

The mapping (8) then reads

$$\tau\left(x_1,\ldots,x_n,\sqrt{1+x^2}\right) = \left(x_1,\ldots,x_n,\sqrt{1+x^2}\right) H(t)$$

with the $(n+1,n+1)$-matrix

$$H(t) = \begin{pmatrix} \cosh t & & & \sinh t \\ & 1 & & \\ & & & \\ & & 1 & \\ \sinh t & & & \cosh t \end{pmatrix}$$

with zeros elsewhere. Let

$$B(p_1,\ldots,p_n;k)$$

be an arbitrary Lorentz boost (see [3, Sections 6.10, 6.11]). We hence have $k \geq 1$,

$$\begin{aligned} &p_1^2+\cdots+p_n^2 < 1, \\ &k^2\left(1-p_1^2-\cdots-p_n^2\right) = 1. \end{aligned} \tag{9}$$

Set $\cosh t := k$, $t \geq 0$, and

$$(a_{11},a_{21},\ldots,a_{n1}) := \frac{\cosh t}{\sinh t}(p_1,\ldots,p_n)$$

for $t > 0$. (For $t = 0$, i.e. $k = 1$, the matrix B must be the identity matrix E, and we are not interested in this case.) Observe

$$a_{11}^2+\cdots+a_{n1}^2 = \frac{k^2}{k^2-1}\sum_{i=1}^{n} p_i^2 = 1$$

from (9). Extend

$$\begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix}$$

to an orthogonal matrix

$$A = \begin{pmatrix} a_{11} & & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

of $\mathbb{R}^n$. Define the so-called *induced Lorentz matrix*

$$\hat{A} := \left(\begin{array}{ccc|c} & & & 0 \\ & A & & \vdots \\ & & & 0 \\ \hline 0 & \dots & 0 & 1 \end{array}\right)$$

and observe

$$B(p_1, \dots, p_n; k) = \hat{A} H(t) \hat{A}^{-1}.$$

(In the case $B = E$ we have $E = EH(0)E^{-1}$.) Because of A.10.1 (see [3. p. 249]), an arbitrary orthochronous Lorentz matrix of $\mathbb{R}^{n+1}$ can be written as the product of a Lorentz boost and an induced Lorentz matrix. This implies that the group $\overset{(n)}{H}$ of all motions of n-dimensional hyperbolic geometry (that is the group of all orthochronous Lorentz matrices of $\mathbb{R}^{n+1}$, see [4, Sections 2.6 and 5.7]), can be generated by $H(t)$, $t \in \mathbb{R}$, and the induced Lorentz matrices, i.e. by linear orthogonal mappings of $\mathbb{R}^n$ and hyperbolic translations concerning the x_1-axis. We now would like to define a function

$$g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$$

as follows: for $\xi \geq 0$ set

$$g(\xi) := d(0, e_1 \sinh \xi).$$

We then have to prove

$$d(x, y) = g(h(x, y))$$

for all $x, y \in \mathbb{R}^n$. Put $h(x, y) =: \xi$. Hence

$$h(x, y) = h(0, e_1 \sinh \xi).$$

Take a linear and orthogonal mapping φ_1 of $\mathbb{R}^n$ that transforms x in $e_1\sqrt{x^2}$, then a τ which maps this latter point into 0. With another φ_2 we get

$$\varphi_2 \tau \varphi_1(x) = 0 \quad \text{and} \quad \varphi_2 \tau \varphi_1(y) =: e_1 \eta$$

with $\eta \geq 0$. Because of

$$\xi = h(x, y) = h(0, e_1 \eta),$$

it follows $\cosh \xi = \cosh h(0, e_1 \eta) = \sqrt{1 + \eta^2}$, i.e.

$$\eta = \sinh \xi.$$

Hence with $\gamma := \varphi_2 \tau \varphi_1$

$$d(x,y) = d(\gamma(x),\gamma(y)) = d(0, e_1 \sinh \xi) = g(\xi) = g(h(x,y)).$$

With respect to the first part of Theorem 2 we know that e and h admit the corresponding mappings mentioned in (D_1) and (D_2). But those mappings already generate the automorphism groups of the geometries in question. ∎

A distance function d of $\mathbb{R}^n$ will be called *additive* on the x_1-axis if, and only if, the following property holds true:

(D_3) *Let α, β, γ be real numbers with $\alpha \leq \beta \leq \gamma$. Then*

$$d(\alpha e_1, \gamma e_1) = d(\alpha e_1, \beta e_1) + d(\beta e_1, \gamma e_1). \tag{10}$$

Theorem 3. *Let d be a distance function of $\mathbb{R}^n$ satisfying (D_1), (D_2), (D_3). Then*

$$d(x,y) = ke(x,y)$$

or

$$d(x,y) = kh(x,y)$$

holds true with a fixed real number $k \geq 0$.

Proof. a) *Euclidean case.* Taking into account Theorem 5 (see [4, Section 5.1]) we only need to prove that (D_3) carries over to every euclidean line of $\mathbb{R}^n$. Let x, z be distinct elements of $\mathbb{R}^n$ and let y be the element

$$y = \lambda x + (1-\lambda) z$$

with $0 \leq \lambda \leq 1$. We then transform x, y, z in

$$\alpha e_1,\ \beta e_1,\ \gamma e_1$$

with $\alpha = 0$, $\beta = (1-\lambda)e(x,z)$, $\gamma = e(x,z)$. Now with Theorem 2

$$d(x,y) = g(e(x,y)) = g(e(0,\beta e_1)) = d(0,\beta e_1)$$

and so on. Hence (10) yields

$$d(x,z) = d(x,y) + d(y,z).$$

Then everything else depends on the solution of the functional equation

$$g(\alpha+\beta) = g(\alpha) + g(\beta)$$

for all $\alpha, \beta \in \mathbb{R}_{\geq 0}$ (see [1]).

b) *Hyperbolic case.* We have to apply Theorem 9 (Section 2.6 in [4]) and a similar procedure as in part a). ∎

Remarks. 1) It is possible now to determine all distance functions d satisfying (D_1), (D_2), constituting a metric. By applying Theorem 2 the reader might verify the next statement which we shall formulate for the hyperbolic case. The situation in question is characterized by all functions

$$g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$$

satisfying

(i) $g(\xi) = 0 \iff \xi = 0$;

(ii) Let α, β, γ be real numbers such that there exists a triangle xyz with $\alpha = h(x,y)$, $\beta = h(y,z)$, $\gamma = h(z,x)$, then

$$g(\gamma) \leq g(\alpha) + g(\beta).$$

2) For general information about hyperbolic geometry compare [5–8].

REFERENCES

1. A c z é l, J. Lectures on functional equations and their applications. Academic Press, New York – London, 1966.
2. A c z é l, J., J. D h o m b r e s. Functional Equations in several variables. New Tork, 1989.
3. B e n z, W. Geometrische Transformationen. BI Wissenschaftsverlag, Mannheim, Leipzig, Wien, Zürich, 1992.
4. B e n z, W. Real Geometries. BI Wissenschaftsverlag, Mannheim, Leipzig, Wien, Zürich, 1994.
5. B e n z, W. Ebene Geometrie. Eine Einführung in Theorie und Anwendungen (to appear).
6. K a r z e l, H., K. S ö r e n s e n, D. W i n d e l b e r g. Einführung in die Geometrie. UTB Vandenhoeck, Göttingen, 1973.
7. N e v a n l i n n a, R., P. K u s t a a n h e i m o. Grundlagen der Geometrie. Basel, 1976.
8. S c h w e r d t f e g e r, H. Geometry of complex numbers. Circle Geometry, Moebius Transformation, Noneuclidean Geometry. Dover Publications, New York, 1979.

Received on 10.06.1996

E-mail address: benz@math.uni-hamburg.de

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

TIKHONOV'S THEOREM FOR FUNCTIONAL-DIFFERENTIAL INCLUSIONS*

TZANKO DONCHEV, IORDAN SLAVOV

We investigate differential inclusions and equations of a retarded type with a small real parameter $\varepsilon > 0$ in part of the derivatives. Analogues of the well-known in the theory of singularly perturbed ordinary differential equations theorem of Thikhonov are obtained. We prove lower semicontinuity of the solution set for inclusions and continuity of the solution for equations in the most appropriate topology when $\varepsilon \to 0$.

Keywords: differential inclusions, equations of retarded type, Tikhonov theorem.
1991/95 Mathematics Subject Classification: 49J40, 49K25, 49J45.

1. INTRODUCTION

Suppose that the functional-differential inclusion

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon \dot{y}(t) \end{pmatrix} \in F(t, x(t), y(t), x_t, y_t), \quad x_0 = \phi, \ y_0 = \psi, \quad t \in I = [0,1], \tag{1}$$

is given, where $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$ and $\varepsilon > 0$ is a real parameter.

In the sequel, $C(I,X)$ and $L^1(I,X)$ are the usual spaces of respectively continuous and integrable functions on I with values in X. For any $z \in C([-\tau,1],\mathbf{R}^k)$ and $t \in I$ we let $z_t \in C([-\tau,0],\mathbf{R}^k)$ be defined by $z_t(s) = z(t+s)$, $-\tau \le s \le 0$.

* Lecture presented at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

This work is partially supported by the National Foundation for Scientific Research at the Bulgarian Ministry of Science and Education, Grant 701/97.

Here $\tau \in (0,1)$ and F is a map from $I \times \mathbf{R}^{n+m} \times C([-\tau,0],\mathbf{R}^n) \times L^1([-\tau,0],\mathbf{R}^m)$ into $\mathbf{R}^n \times \mathbf{R}^m$, while $\phi \in C([-\tau,0],\mathbf{R}^n)$, $\psi \in C([-\tau,0],\mathbf{R}^m)$.

There is a fundamental theorem refered as Tikhonov's theorem [10] dealing with the continuity of the (unique) solution of (1) when F is single valued and does not contain (x_t, y_t). Namely, continuous dependence of the solution with respect to $C(I,\mathbf{R}^n) \times C([\delta,1],\mathbf{R}^m)$ topology $(0 < \delta < 1)$ when $\varepsilon \to 0$ is stated. Our considerations differ from the situation in [10] also in the fact that we assume only measurable on t right hand side. Then it is natural to define the solution set $Z(\varepsilon)$ of (1) when $\varepsilon > 0$ as the collection of all absolutely continuous functions (x,y) satisfying (1) for a.e. $t \in I$. When $\varepsilon = 0$, inclusion (1) becomes

$$\begin{pmatrix} \dot{x}(t) \\ 0 \end{pmatrix} \in F(t, x(t), y(t), x_t, y_t), \quad x_0 = \phi,\ y_0 = \psi, \quad t \in I = [0,1]. \tag{2}$$

Here solutions are all pairs (x,y) of absolutely continuous functions $x(\cdot)$ and L^1-functions $y(\cdot)$ such that (2) holds for a.e. $t \in I$. As in the ordinary differential case, $y(\cdot)$ can differ from the initial condition $\psi(\cdot)$ at $t = 0$.

It is too restrictive to assume the y-part of the solutions of the above "degenerate" inclusion to be continuous in view of the following simple example:

$$\varepsilon \dot{y}(t) = -2ay(t) + ay\left(t - \frac{1}{2}\right), \ \ y(s) = 1, \ \ s \in \left[-\frac{1}{2}, 0\right), \ \ a > 0.$$

For $\varepsilon = 0$ one has $0 = -2y(t) + y(t - 1/2)$, i.e. $y(t) = (1/2)y(t - 1/2)$. Thus $y^0(t) = 1/2$ for $t \in [0, 1/2)$ and $y^0(t) = 1/4$ for $t \in [1/2, 1)$. For this reason the C-topology used in [10] is not suitable and must be replaced with another one. In Examples 2.1 and 2.2 we show that when the delay is not fixed it happens the classical Tikhonov's theorem not to be valid. So it must be reformulated in the functional-differential case when it holds at all.

Here we examine first the lower semicontinuity properties of the solution map $Z(\varepsilon)$ as $\varepsilon \to 0^+$ and then derive on this base the continuity dependence of the solution for equations. For inclusions without the functional arguments (x_t, y_t) the lower semicontinuity is studied initially in [11]. The results then are extended under weaker type of assumptions in [3] for functional-differential inclusions with fixed time delay. The main assumption in the last paper is a version of the one-side Lipschitz condition used first for multivalued maps in [2]. Since singular perturbations are not presented in [2], this key condition is modified in [3] and here in a suitable way. We do not consider upper semicontinuous properties since, as shown in [3], the solution set is not upper semicontinuous in used here $C(I,\mathbf{R}^n) \times L^1(I,\mathbf{R}^m)$ topology, even for linear control system. Moreover, in the case considered in [3], redefining the solution set of (2) to obtain upper semicontinuity one will lose lower semicontinuity. Some upper semicontinuous results under restrictive assumptions are obtained in [3–5].

At the end of the section we shall give some notations and definitions. Introduce the subspaces $\Omega_i = \{\alpha \in C([-\tau,0],\mathbf{R}^{k_i}) : |\alpha(0)| = \max_{-\tau \le s \le 0} |\alpha(s)|\}$, $k_1 = n$, $k_2 = m$, which are used in Razumikhin type conditions [7]. The norms in $C(I,X)$ and $L^1(I,X)$ are denoted with $\|\cdot\|_C$ and $\|\cdot\|_{L^1}$, respectively. For the

sake of simplicity we will denote by $\|\alpha_t\|_C$ and $\|\alpha_t\|_{L^1}$, respectively, the norms $\max_{-\tau\le s\le 0}|\alpha(t+s)|$ and $\int_{-\tau}^{0}|\alpha(t+s)|\,ds$. For a set $A\subset\mathbf{R}^k$ and a vector $l\in\mathbf{R}^k$ we let $\sigma(l,A)=\sup_{a\in A}\langle l,a\rangle$ be the support function, where $\langle\cdot,\cdot\rangle$ is the scalar product. If $A\subset\mathbf{R}^{n+m}$, we denote by $\hat{A}$ the projection of A on $\mathbf{R}^n$, by $\bar{A}$ the projection of A on $\mathbf{R}^m$, and by clA ($clcoA$) the closed (the closed convex) hull of A. The set-valued map $G: I\times Z\to Z$ is called: a) lower semicontinuous (LSC) when for every (t,z) and every $u\in G(t,z)$ there exists $u_i\in G(t_i,z_i)$ such that $u_i\to u$ when $t_i\to t$, $z_i\to z$; b) upper semicontinuous (USC) if for every (t,z) and every $\nu>0$ there exists $\delta>0$ such that $G(s,w)\subset G(t,z)+\nu U$ (here U is the unit ball in Z) when $|t-s|+|z-w|<\delta$; c) continuous when G is LSC and USC. G is called *almost* continuous (resp. LSC, USC) when for every $\delta>0$ there is a compact set $I_\delta\subset I$ with $meas(I\setminus I_\delta)<\delta$ such that G is continuous (resp. LSC, USC) on $I_\delta\times Z$. For more detailed considerations of definitions and concepts used bellow we refer to [1] and [7].

2. LOWER SEMICONTINUITY IN $C\times L^1$–TOPOLOGY

We take an example which tells us that for continuity with respect to $C[\delta,1]$ topology on $y(\cdot)$ there have to be restrictive assumptions.

Example 2.1. Consider the following equation:

$$\varepsilon\dot{y}(t)=-2y(t)+\max_{s\in I_t}y(t+s),\quad y(0)=1,$$

where $I_t=[\max\{-1/2,-t\},0]$ for $t\in[0,1]$. For $\varepsilon>0$ one can find

$$y^\varepsilon(t)\ge\frac{1}{2}\left(1+\exp\left(-\frac{1}{\varepsilon}\right)\right),\quad 0\le t\le\frac{1}{2},$$

$$y^\varepsilon(t)\ge\frac{1}{4}\left(1+\exp\left(-\frac{2}{\varepsilon}\left(t-\frac{1}{2}\right)\right)\right),\quad \frac{1}{2}\le t\le 1.$$

For $\varepsilon=0$ we get the "degenerate" equation

$$2y(t)=\max_{s\in I_t}y(t+s).$$

Obviously, $\bar{y}^0(t)=1/2$, $t\in(0,1/2]$; $\bar{y}^0(t)=1/4$, $t\in(1/2,1]$ with $\bar{y}^0(0)=1$ is a solution of the above equation. Also it is not difficult to see that $y^\varepsilon(t)\to\bar{y}^0(t)$, $\varepsilon\to 0$ for $t\in I$ and that this convergence is uniform on $[\delta,1/2)\cup[1/2+\delta,1]$. On the other hand, $y^0(t)\equiv 0$ on $t\in I$ is other solution of the "degenerate" equation. The last implies that there is no continuous in $C[\delta,1]$ but only USC dependence in $C([\delta,1/2)\cup[1/2+\delta,1])$ topology.

Example 2.2. Let us combine the above equation with the control system from Example 2.5 of [3], i.e. consider

$$\dot{x}=|y_1-2y_2|,\quad x(0)=0,$$

$$\varepsilon\dot{y}_1 = -y_1 + u(t), \quad y_1(0) = 0,$$
$$\varepsilon\dot{y}_2 = -2y_2 + u(t), \quad y_2(0) = 0,$$
$$\varepsilon\dot{y}_3 = -2y_3(t) + \max_{s\in I_t} y_3(t+s), \quad y_3(0) = 1,$$

where $u(t) \in [-1,1]$ is measurable. It is shown in [3] that the solution set of the subsystem consisting of the first three equations is not USC in $C([0,1],\mathbf{R}) \times L^1([0,1],\mathbf{R}^2)$ topology at $\varepsilon = 0$. Thus the solution set of the above inclusion is neither LSC nor USC.

These examples tell us that when the delay depends on time t it is hard to expect that Tikhonov's theorem is true. But still there are situations in which we could formulate a very close result. Consider first (1) under the following assumptions:

A1. The map F is almost continuous and bounded on the bounded sets. Moreover, there exist constants $a, b, \mu > 0$ such that for every $(x,y) \in \mathbf{R}^{n+m}$

$$\sigma(x, \hat{F}(t,x,y,\alpha,\beta)) \le a(1 + |x|^2 + |y|^2 + \|\beta\|_C^2), \quad \alpha \in \Omega_1,\ \beta \in C([-\tau,0],\mathbf{R}^m),$$
$$\sigma(y, \bar{F}(t,x,y,\alpha,\beta)) \le b(1 + |x|^2 + \|\alpha\|_C^2) - \mu|y|^2, \quad \alpha \in C([-\tau,0],\mathbf{R}^n),\ \beta \in \Omega_2,$$

for a.e. $t \in I$. Here $\alpha(0) = x$, $\beta(0) = y$.

A2. There exist positive constants A, B and μ such that if we choose arbitrary $(x_i, y_i, \alpha_i, \beta_i) \in \mathbf{R}^{n+m} \times C([-\tau,0],\mathbf{R}^n) \times L^1([-\tau,0],\mathbf{R}^m)$, $i = 1,2$, then for every $(f_1,g_1) \in F(t,x_1,y_1,\alpha_1,\beta_1)$ there is $(f_2,g_2) \in F(t,x_2,y_2,\alpha_2,\beta_2)$ such that

$$\langle x_1 - x_2, f_1 - f_2\rangle \le A(|x_1 - x_2|^2 + |y_1 - y_2|^2 + \|\beta_1 - \beta_2\|_{L^1}^2), \text{ for } \alpha_1 - \alpha_2 \in \Omega_1,$$
$$\langle y_1 - y_2, g_1 - g_2\rangle \le B(|x_1 - x_2|^2 + \|\alpha_1 - \alpha_2\|_C^2 + \|\beta_1 - \beta_2\|_{L^1}^2) - \mu|y_1 - y_2|^2$$

for a.e. $t \in I$. Here $\alpha_i(0) = x_i$ and for β_i continuous $\beta_i(0) = y_i$, $i = 1,2$.

The next result is proved in [3].

Lemma 2.3. *Under* A1 *there exists a constant* $M > 0$ *such that* $|x^\varepsilon(t)| + |y^\varepsilon(t)| \le M$ *for every* $t \in I$, $(x^\varepsilon, y^\varepsilon) \in Z(\varepsilon)$ *and* $\varepsilon > 0$, *and a.e. on* I *if* $\varepsilon = 0$.

By A1 it follows that there exists $L > 0$ such that $|F(t,x,y,\alpha,\beta)| \le L$ for every $t \in I$, $|x| + |y| \le M + 1$ and $\|\alpha\|_C + \|\beta\|_{L^\infty} \le M + 1$.

Theorem 2.4. *Under assumptions* A1 *and* A2 *the solution set* $Z(\varepsilon)$ *is LSC at* $\varepsilon = 0^+$ *with respect to* $C([0,1],\mathbf{R}^n) \times L^1([0,1],\mathbf{R}^m)$ *topology.*

Proof. Let (x^0, y^0) be a solution of (2) and $\delta > 0$ be given. Then there is a Lipschitz on I function z with a Lipschitz constant K_δ such that $z(s) = \psi(s)$, $s \in [-\tau, 0]$, and

$$\|z - y^0\|_{L^1} \le \delta, \quad \|\rho\|_{L^1} \le \delta.$$

Here $\rho(t) = D_H(F(t,x^0,y^0,x_t^0,y_t^0), F(t,x^0,z,x_t^0,z_t))$ and $D_H(\cdot,\cdot)$ is the Hausdorff distance between sets. Therefore

$$d((\dot{x}^0(t), \varepsilon\dot{z}(t)), F(t,x^0(t),z(t),x_t^0,z_t)) \le \varepsilon K_\delta + \rho(t). \tag{3}$$

Introduce the following conditions:

$$\begin{aligned} &\langle x^0(t)-u, \dot{x}^0(t)-f\rangle \\ &< 2A(|x^0(t)-u|^2+|z(t)-v|^2+\|z_t-\beta\|^2_{L^1})+\varepsilon K_\delta+\rho(t)+\delta, \qquad (4) \\ &\langle z(t)-v, \varepsilon\dot{z}(t)-g\rangle \\ &< 2B(|x^0(t)-u|^2+\|x^0_t-\alpha\|^2_C+\|z_t-\beta\|^2_{L^1})-\mu|z(t)-v|^2+\varepsilon K_\delta+\rho(t)+\delta. \qquad (5) \end{aligned}$$

Consider the map $\Gamma_\delta(t,u,v,\alpha,\beta)$, which we define only for continuous β, with values as follows:

a) $cl\{(f,g)\in F(t,u,v,\alpha,\beta) : g$ satisfies (5)$\}$ for $\alpha - x^0_t \notin \Omega_1, u=\alpha(0)$ and $v=\beta(0)$;

b) $cl\{(f,g)\in F(t,u,v,\alpha,\beta) : (f,g)$ satisfies (4) and (5)$\}$ for $\alpha - x^0_t \in \Omega_1, u=\alpha(0)$ and $v=\beta(0)$;

c) $\Gamma_\delta(t,u,v,\alpha,\beta)=F(t,u,v,\alpha,\beta)$ when $u\neq\alpha(0)$ or $v\neq\beta(0)$.

Note that F is almost continuous on $I\times\mathbf{R}^{n+m}\times C(I,\mathbf{R}^{n+m})$. We claim that $\Gamma_\delta(\cdot)$ is almost LSC with nonempty and compact values. To prove that we first note that $\Gamma_\delta(\cdot)$ is compact valued by its definition, Lemma 2.3 and A1. We will show the nonemptiness of $\Gamma_\delta(\cdot)$ only in case b).

By (3) there is $(f^0(t),g^0(t))\in F(t,x^0(t),z(t),x^0_t,z_t)$ such that for a.e. $t\in I$

$$|(\dot{x}^0(t),\varepsilon\dot{z}(t))-(f^0(t),g^0(t))|\leq\varepsilon K_\delta+\rho(t).$$

So, there exists $(f,g)\in F(t,u,v,\alpha,\beta)$ such that for $x_1=x^0(t)$, $x_2=u$, $y_1=z(t)$, $y_2=v$, $f_1=f^0$, $f_2=f$, $g_1=g^0$ and $g_2=g$ the inequalities of A2 hold, i.e.

$$\begin{aligned} \langle x^0(t)-u, f^0-f\rangle &< A(|x^0(t)-u|^2+|z(t)-v|^2+\|z_t-\beta\|^2_{L^1}), \\ \langle z(t)-v, g^0-g\rangle &< B(|x^0(t)-u|^2+\|x^0_t-\alpha\|^2_C+\|z_t-\beta\|^2_{L^1})-\mu|z(t)-v|^2. \end{aligned}$$

Therefore the inequalities (4) and (5) are fulfilled.

The fact that $\Gamma_\delta(\cdot)$ is almost LSC has a standard proof (see [1]), which is omitted.

Now, from [6] we know that the inclusion

$$\begin{pmatrix}\dot{x}(t)\\ \varepsilon\dot{y}(t)\end{pmatrix}\in\Gamma_\delta(t,x(t),y(t),x_t,y_t), \quad x_0=\phi, y_0=\psi, \quad t\in I=[0,1], \qquad (6)$$

has a solution $(x^\varepsilon,y^\varepsilon)$ in this case as well. On the other hand, $|x^\varepsilon(t)-x^0(t)|^2\leq 2h(t)$ and $|y^\varepsilon(t)-z(t)|^2\leq 2r(t)$, where:

$$\begin{aligned} \dot{h}(t) &= 2A(h(t)+r(t)+\|r_t\|_{L^1})+\rho(t)+\delta+\varepsilon K_\delta, \quad h(0)=0, \\ \varepsilon\dot{r}(t) &= 2Bh(t)-\mu r(t)+2B(\|h_t\|_C+\|r_t\|_{L^1})+\rho(t)+\delta+\varepsilon K_\delta, \quad r(0)=r_0. \end{aligned}$$

We do not indicate the dependence on ε of the solution of the system for the sake of simplicity of notations. Let k be a sufficiently large natural number. We divide $[0,1]$ on k parts with equal lengths. Obviously, by the first equation above we have that $h(\cdot)$ increases, i.e one can suppose without loss of generality that $h(t)=\|h_t\|_C$.

Then solving the second equation on $[0, 1/k]$ and integrating by parts one obtains

$$\begin{aligned} r(t) &\le \exp(-\mu t/\varepsilon) r_0 + (1/\varepsilon) \int_0^t \exp(-\mu(t-s)/\varepsilon)(4Bh(s) + \rho(s) \\ &\quad + 2B\|r_s\|_{L^1} + \delta + \varepsilon K_\delta)\, ds \\ &\le \exp(-\mu t/\varepsilon) r_0 + (1/\varepsilon) \int_0^t \exp(-\mu(t-s)/\varepsilon)(\rho(s) + 2B\|r_s\|_{L^1})\, ds \\ &\quad + (1/\mu)(4Bh(t) + \delta + \varepsilon K_\delta). \end{aligned}$$

Denoting further with C an arbitrary positive constant dependent only on A, B and μ (in the following inequality for example $C = 2A + 8AB/\mu$), we derive that

$$h(t) \le \int_0^t \exp(C(t-s))(\rho(s) + 2A\|r_s\|_{L^1} + (1 + 2A/\mu)(\delta + \varepsilon K_\delta) + 2A \exp(-\mu s/\varepsilon) r_0)\, ds$$

$$+ (2A/\varepsilon) \int_0^t \int_0^s \exp(C(t-s)) \exp(-\mu(s-\lambda)/\varepsilon)(\rho(\lambda) + \|r_\lambda\|_{L^1})\, d\lambda\, ds.$$

Thus changing the order of integration we get $h(t) \le C(2\varepsilon K_\delta + \delta + 1/k)$ for $t \in [0, 1/k]$. Consequently,

$$\int_0^t |r(s)|\, ds \le C\left(2\varepsilon K_\delta + \delta + \frac{1}{k}\right) \quad \text{for } t \in \left[0, \frac{1}{k}\right].$$

By induction one can show that

$$h(t) \le C\left(2\varepsilon K_\delta + \delta + \frac{1}{k} + \frac{1}{k^2} + \cdots + \frac{1}{k^i}\right),$$

$$\|r\|_{L^1[0,t]} \le C\left(2\varepsilon K_\delta + \delta + \frac{1}{k} + \frac{1}{k^2} + \cdots + \frac{1}{k^i}\right), \quad t \in \left[0, \frac{i}{k}\right].$$

Finally, one obtains

$$h(t) \le C\left(2\varepsilon K_\delta + \delta + \frac{1}{k-1}\right), \quad \|r(\cdot)\|_{L^1} \le C\left(2\varepsilon K_\delta + \delta + \frac{1}{k-1}\right).$$

Since k is arbitrarily large, we get that there exists a solution $(x^\varepsilon, y^\varepsilon)$ of (1) such that

$$\|x^\varepsilon - x^0\|_C^2 \le C(\varepsilon K_\delta + \delta), \quad \|y^\varepsilon - y^0\|_{L^1}^2 \le C(\varepsilon K_\delta + \delta).$$

Since δ is arbitrary and K_δ depends on δ but not on ε, the LSC in the considered topology is established. ■

Remark. A preliminary version of this theorem is reported in [9].

Consider the following special case of (1):

$$\begin{pmatrix} \dot{x}(t) \\ \varepsilon\dot{y}(t) \end{pmatrix} \in F(t, x(t), y(t), x(t-\tau(t)), y(t-\tau(t))), \tag{7}$$

$$x(t) = \phi(t),\ y(t) = \psi(t), \quad t \in [-\lambda, 0],$$

where $\tau(t) \in [0, \lambda]$ is a monotone non-increasing function on I. Suppose that:

B1. The map F is Caratheodory's and bounded on the bounded sets. Moreover, there exist constants $a, b, \mu > 0$ such that for every $t \in I$, $(x(t), y(t)) \in \mathbf{R}^{n+m}$

$$\sigma(x(t), \hat{F}(t, x(t), y(t), x(t-\tau(t)), y(t-\tau(t)))) \le a(1 + |x(t)|^2 + |y(t)|^2 + |x(t-\tau(t))|^2 + |y(t-\tau(t))|^2),$$

$$\sigma(y(t), \bar{F}(t, x(t), y(t), x(t-\tau(t)), y(t-\tau(t)))) \le b(1 + |x(t)|^2 + |x(t-\tau(t))|^2 + |y(t-\tau(t))|^2) - \mu|y(t)|^2.$$

B2 (one-side Lipschitz condition). There exist positive constants A, B and μ such that for every $(f_1, g_1) \in F(t, x_1(t), y_1(t), x_1(t-\tau(t)), y_1(t-\tau(t)))$ there is $(f_2, g_2) \in F(t, x_2(t), y_2(t), x_2(t-\tau(t)), y_2(t-\tau(t)))$ such that

$$\langle x_1 - x_2, f_1 - f_2\rangle \le A(|x_1 - x_2|^2 + |y_1 - y_2|^2 + |\alpha_1 - \alpha_2|^2 + |\beta_1 - \beta_2|^2),$$

$$\langle y_1 - y_2, g_1 - g_2\rangle \le B(|x_1 - x_2|^2 + |\alpha_1 - \alpha_2|^2 + |\beta_1 - \beta_2|^2) - \mu|y_1 - y_2|^2$$

for a.e. $t \in I$. Here $\alpha_i(t) = x_i(t-\tau(t))$, $\beta_i(t) = y_i(t-\tau(t))$, $i = 1, 2$.

B3. If $\inf_{t\in I} \tau(t) = 0$, then $\mu > B$.

Theorem 2.5. *Under the assumptions* B1–B3, *the solution set* $Z(.)$ *is lower semicontinuous in* $C(I, R^n) \times L^1(I, R^m)$ *topology.*

Proof. Define the sequence $t_{i+1} = \sup\{t \in I | t_{i-1} \le t - \tau(t) \le t_i\}$, where $t_0 = -\lambda$, $t_1 = 0$. There are two cases. If $t_k = 1$ for some k, one can easily complete the proof exploiting the same fashion as in a fixed time lag, see Theorem 3.2 from [3]. In the opposite case there exists obviously $\nu \le 1$ with $\nu = \lim_{i\to\infty} t_i$. Then **B3** holds. Moreover, $\tau(t) = 0$ for $t \ge \nu$, i.e. the inclusion (7) becomes an ordinary differential one. Let $\delta > 0$ be given and (x^0, y^0) be the solution of (7) for $\varepsilon = 0$. Then for every $t < \nu$ again on the base of [3] one can find $\varepsilon(t, \delta)$ such that there exists $(x^\varepsilon, y^\varepsilon) \in Z(\varepsilon)$ whenever $0 \le \varepsilon \le \varepsilon(t, \delta)$ with

$$\|x^0 - x^\varepsilon\|_{C[0,t]} + \|y^0 - y^\varepsilon\|_{L^1[0,t]} < \delta/3.$$

Note that the norms above are evaluated on $[0, t]$. Moreover, $Z(\varepsilon)$ is LSC on $[\nu, 1]$ with respect to $C([\nu, 1], \mathbf{R}^n) \times L^1([\nu, 1], \mathbf{R}^m)$, see [11]. So without loss of generality one can suppose that

$$\|x^0 - x^\varepsilon\|_{C[\nu,1]} + \|y^0 - y^\varepsilon\|_{L^1[\nu,1]} < \delta/3.$$

Using the boundedness of the solution set and thus of the right hand side of (7), we can manage also on the interval $[t, \nu]$. Namely, if $\nu - t$ is small enough, then

$$\|x^0 - x^\varepsilon\|_{C[t,\nu]} + \|y^0 - y^\varepsilon\|_{L^1[t,\nu]} < \delta/3.$$

Consequently, there exists $(x^\varepsilon, y^\varepsilon) \in Z(\varepsilon)$ such that

$$\|x^0 - x^\varepsilon\|_C + \|y^0 - y^\varepsilon\|_{L^1} < \delta$$

for sufficiently small ε. ■

3. TIKHONOV TYPE THEOREM FOR FUNCTIONAL-DIFFERENTIAL EQUATIONS

Consider now the following singularly perturbed system of functional-differential equations:

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), y(t), x_t, y_t), \quad x_0 = \phi, \\ \varepsilon\dot{y}(t) &= g(t, x(t), y(t), x_t, y_t), \quad y_0 = \psi, \end{aligned} \tag{8}$$

derived from (1) when F is single valued. Here $f(\cdot)$ and $g(\cdot)$ are Caratheodory's functions, satisfying A1 and A2.

First we shall show that the reduced system

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), y(t), x_t, y_t), \quad x_0 = \phi, \\ 0 &= g(t, x(t), y(t), x_t, y_t), \quad y_0 = \psi, \end{aligned} \tag{9}$$

admits $C(I, \mathbf{R}^n) \times L^1(I, \mathbf{R}^n)$ solution, i.e. the next lemma is true.

Lemma 3.1. *Under the assumptions* A1 *and* A2 *the degenerate system* (9) *has a solution.*

Proof. First we shall consider the case when f and g are jointly continuous, i.e. continuous in all arguments.

By Lemma 2.3 for $0 < \delta < \mu$ there is a constant M_δ such that for all $t \in I$

$$\begin{aligned} &|x(t)| + |y(t)| \le M_\delta, \text{ when} \\ &|\dot{x}(t) - f(t, x(t), y(t), x_t, y_t)| \le \delta, \quad |g(t, x(t), y(t), x_t, y_t)| \le \delta. \end{aligned}$$

Choose a sequence $\delta_i \to 0^+$ and construct the corresponding sequence of approximate solutions (x^i, y^i) as follows. By the well-known theorem of Minty-Browder there exists $\beta_0 \in \mathbf{R}^m$ such that

$$0 = g(t, \phi(0), \beta_0, \phi, \psi). \tag{10}$$

Let

$$x^i(t) = \phi(0) + tf(0, \phi(0), \beta_0, \phi, \psi), \quad y^i(t) = \beta_0,$$

for $t \in [0, \nu_1]$. Here ν_1 is the maximal ν for which (10) and

$$|\dot{x}^i(t) - f(t, x^i(t), y^i(t), x^i_t, y^i_t)| \le \delta_i, \quad |g(t, x^i(t), y^i(t), x^i_t, y^i_t)| \le \delta_i$$

hold on $[0, \nu]$. Using continuity of f, g and Zorn's lemma, it is not difficult to show the existence of such (x^i, y^i) on the whole I. By the Arzela-Ascoli's theorem $\{x^i(\cdot)\}_{i=1}^\infty$ is $C(I, \mathbf{R}^n)$ precompact and passing to subsequences if needed, there exists a cluster point $x^0(\cdot) \in C(I, \mathbf{R}^n)$. We shall show that $\{y^i(\cdot)\}_{i=1}^\infty$ is a Cauchy

sequence in $L^1(I, \mathbf{R}^m)$. Denote $r(t) \equiv r_{ij}(t) = |y^i(t) - y^j(t)|$, $\delta_{ij} = \|x^i(\cdot) - x^j(\cdot)\|_C$. Then, of course, $\|x_t^i - x_t^j\|_{L^1} \le \delta_{ij}$ and by A2 we obtain

$$\mu r^2(t) \le B(\delta_{ij}^2 + \|r_t\|_{L^1}^2) + C(\delta_i + \delta_j).$$

For the sake of simplicity of notations here and further we denote with C an arbitrary constant and with δ_{ij} an expression tending to zero with $i, j \to \infty$. Hence

$$r(t) \le C(\delta_{ij} + \|r_t\|_{L^1}), \ t \in I \text{ and } \ r(t) = 0, \ t \in [-\tau, 0].$$

Let $r(t) = M$ on $[0, \tau]$. Thus $\|r_t\|_{L^1} \le \int_0^t M \, ds = Mt$ for $t \in [0, \tau]$. Therefore

$$r(t) \le C\delta_{ij} + CMt, \ t \in [0, \tau].$$

Since $\|r_t\|_{L^1} = \int_0^\tau r(t - s) \, ds = \int_0^t r(s) \, ds$, we have

$$\|r_t\|_{L^1} \le C\delta_{ij} t + CM\frac{t^2}{2!}, \ t \in [0, \tau].$$

Then it follows

$$r(t) \le C\delta_{ij}\left(1 + \frac{Ct}{1!}\right) + CM\frac{t^2}{2!}, \ t \in [0, \tau].$$

Proceeding in the same way, we find that

$$r(t) \le C\delta_{ij}\left(1 + \frac{Ct}{1!} + \frac{(Ct)^2}{2!} + \cdots\right) + M \lim_{n\to\infty} \frac{(Ct)^n}{n!} \le \delta_{ij} C \exp(Ct), \ t \in [0, \tau].$$

Thus $\lim_{i,j\to\infty} r(t) \equiv \lim_{i,j\to\infty} r_{ij}(t) = 0$ and $\{y^i(\cdot)\}_{i=1}^\infty$ is a Cauchy sequence on $[0, \tau]$. Therefore $\lim_{i\to\infty} y^i(t) = y(t), t \in [0, \tau]$ exists. It is easy to show that $(x(t), y(t))$ is a solution of (9) on $[0, \tau]$. Analogously (keeping in mind that $r(t) = 0, t \in [0, \tau]$),the solution can be extended on $[\tau, 2\tau]$ and therefore by induction on $[0, 1]$.

Now let $f(\cdot)$ and $g(\cdot)$ be Caratheodory's functions. By Scorza-Dragoni's theorem $f(\cdot)$ and $g(\cdot)$ are almost continuous, so we can use the same fashion. Namely, for $\delta_i > 0$ consider $A_i \subset I$ with $meas A_i < \delta_i$, $A_{i+1} \subset A_i$. Also let us have on $I \setminus A_i$ that $f(\cdot)$ and $g(\cdot)$ are continuous and for the approximate solutions (x^i, y^i) the following relations are true:

$$\left|\dot{x}^i(t) - f(t, x^i(t), y^i(t), x_t^i, y_t^i)\right| \le \delta_i, \quad \left|g(t, x^i(t), y^i(t), x_t^i, y_t^i)\right| \le \delta_i.$$

On A_i the above distances are less or equal to L.

Denote again $r(t) = |y^i - y^j|$. One can show that $r(t) \le \delta_{ij}(t) D \exp(t)$, where $\delta_{ij}(t) \le M, t \in A_i$, and $\delta_{ij}(t) \le \delta_{ij}, t \in I \setminus A_i$, where $\lim_{i,j\to\infty} \delta_{ij} = 0$. Therefore $(x^i(\cdot), y^i(\cdot)) \to (x(\cdot), y(\cdot))$, which is a solution of (9) on $[0, 1]$. ∎

Now one can easily prove the next variant of the Tikhonov's theorem.

Theorem 3.2. *Under conditions* A1, A2 *for single valued* F *the solution set* $Z(\varepsilon)$ *of* (8) *is continuous in* $C([0, 1], \mathbf{R}^n) \times L^1([0, 1], \mathbf{R}^m)$ *topology at* $\varepsilon = 0^+$.

Proof. The solution set $Z(0)$ of (9) is non-empty thanks to Lemma 3.1. By A2 it follows (see [8]) that $Z(\varepsilon)$ is single valued. Then by the LSC of $Z(\varepsilon)$ at $\varepsilon = 0^+$ (Theorem 2.4) the proof is completed. ■

REFERENCES

1. Deimling, K. Multivalued differential equations. Walter de Gruyter, Berlin, 1992.
2. Donchev, T. Functional differential inclusions with monotone right-hand side. — Nonlinear Analysis TMA, **16**, 1991, 533–542.
3. Donchev, T., I. Slavov. Singularly perturbed functional-differential inclusions. — Set-Valued Analysis, **3**, 1995, 113–128.
4. Dontchev, A., I. Slavov. Upper semicontinuity of solutions of singularly perturbed differential inclusions. — System Modelling and Optimization, Lecture Notes in Control and Inf. Sc., **143**, Springer 1991, 273–280.
5. Dontchev, A., Tz. Donchev, I. Slavov. A Tikhonov-type theorem for singularly perturbed differential inclusions. — Nonlinear Analysis TMA, **26**, 1996, 1547–1554.
6. Fryszkowski, A. Existence of solutions of functional-differential inclusions in nonconvex case. — Ann. Polon. Math., **45**, 1989, 121–124.
7. Hale, J. Theory of functional differential equations. Springer, 1977.
8. Lakshmikantham, V., A. Mitchel, R. Mitchel. On the Existance of Solutions of Differential Equations of Retarded Type in Banach Space. — Ann. Polon. Math., **35**, 1977, 253–260.
9. Slavov, I., T. Donchev. Singularly perturbed functional-differential inclusions — application to optimal control. In: 20th Summer School "Applications of Mathematics in Engineering, Varna'94", 1994, 135–142.
10. Tikhonov, A. Systems of differential equations containing a small parameter in the derivatives. — Mat. Sbornik, **31** (**73**), 1952, 575–586 (in Russian).
11. Veliov, V. Differential inclusions with stable subinclusions. — Nonlinear Analysis, **21**, 1994, 1027–1038.

Received on 9.07.1996
Revised on 18.11.1997

Tz. Donchev
Dept. of Mathematics
University of Mining and Geology
1100 Sofia, Bulgaria
e-mail: donchev@cserv.mgu.bg

Iordan Slavov
Inst. Appl. Math. & Inf., bl.2
Technical University
1000 Sofia, Bulgaria
e-mail: iis@vmei.acad.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

COMPLETE SYSTEMS OF BESSEL AND INVERSED BESSEL POLYNOMIALS IN SPACES OF HOLOMORPHIC FUNCTIONS*

JORDANKA PANEVA-KONOVSKA

Let $B_n(z)$, $n = 0, 1, \ldots$, be the Bessel polynomials generated by

$$(1-4zw)^{-1/2} \exp\left\{\frac{1-(1-4zw)^{1/2}}{2z}\right\} = \sum_{n=0}^{\infty} B_n(z) w^n, \quad |4zw| < 1$$

and the functions $\widetilde{B}_n(z)$ be defined by the relations

$$\widetilde{B}_n(z) = 4^{-n} z^n B_n(1/z) \exp(-z/2).$$

Let $K = \{k_n\}_{n=0}^{\infty}$ be an increasing sequence of non-negative integers.

Sufficient conditions for the completeness of the systems $\{B_{k_n}(z)\}_{n=0}^{\infty}$ and $\left\{\widetilde{B}_{k_n}(z)\right\}_{n=0}^{\infty}$ in spaces of holomorphic functions are given in terms of the density of the sequence K.

Keywords: holomorphic functions, complete systems, Bessel polynomials.
Mathematics Subject Classification: 30B60, 33D25, 41A58.

1. INTRODUCTION

Let G be an arbitrary region in the complex plane $\mathbb{C}$ and $H(G)$ be the space of the complex functions holomorphic in G. As usual, we consider $H(G)$ with the topology of uniform convergence on compact subsets of G. A system $\{\varphi_n(z)\}_{n=0}^{\infty} \subset$

* Lecture presented at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

This work was partially supported by the Ministry of Education and Science, Bulgaria, under Project MM 433/94.

$H(G)$ is called complete in $H(G)$ if for every $f \in H(G)$, every compact set $K \subset G$ and every $\varepsilon > 0$ there exists a linear combination

$$P(z) = \sum_{n=0}^{N} c_n \varphi_n(z), \quad c_n \in \mathbb{C}; \quad n = 0, 1, 2, \ldots, N,$$

such that $|f(z) - P(z)| < \varepsilon$ whenever $z \in K$. For example, if $G \subset \mathbb{C}$ is simply connected, the system $\{z^n\}_{n=0}^{\infty}$ is complete in $H(G)$ and this assertion is nothing but a particular case of the Runge's approximation theorem [1, (2.1), p. 176].

Let γ be a Jordan curve in $\mathbb{C}$ and C_γ be the closure of its outside with respect to the extended complex plane $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. By H_γ we denote the (vector) space of all complex functions, holomorphic in an open set containing C_γ and vanishing at infinity. The next statement is a criterion for completeness in the space $H(G)$ [2, Theorem 17, p. 211].

(CC) A system $\{\varphi_n(z)\}_{n=0}^{\infty}$ of complex functions holomorphic in a simply connected region $G \subset \mathbb{C}$ is complete in the space $H(G)$ iff for every rectifiable Jordan curve $\gamma \subset G$ and every function $F \in H$ the equalities

$$\int_\gamma F(z)\varphi_n(z)\,dz = 0, \quad n = 0, 1, 2, \ldots,$$

imply $F \equiv 0$.

Completeness of systems of special functions in spaces of holomorphic functions has been considered also by Kazmin [3], Leontiev [4, Ch. 3], Rusev [5–9].

2. BESSEL AND INVERSED BESSEL POLYNOMIALS

Let us define the function $\Phi(z, w)$ as

$$\Phi(z, w) = (1 - 4zw)^{-1/2} \exp\left\{\frac{1 - (1 - 4zw)^{1/2}}{2z}\right\}, \quad |4zw| < 1. \tag{2.1}$$

Note that the identity

$$\frac{1 - (1 - 4zw)^{1/2}}{2z} = \frac{2w}{1 + (1 - 4zw)^{1/2}} \tag{2.2}$$

implies that the point $z = 0$ is a removable singularity of this function for every fixed w.

Let $B_n(z)$, $n = 0, 1, \ldots$, be the Bessel polynomials defined by [10, (11.2), VII]

$$\Phi(z, w) = \sum_{n=0}^{\infty} B_n(z) w^n, \quad |4zw| < 1. \tag{2.3}$$

The polynomials $y_n(x; a, b)$ [11, 6] are defined by

$$(1 - 2xt)^{-1/2} \exp\left(\frac{1}{2} - \frac{1}{2}(1 - 2xt)^{1/2}\right)^{2-a} \exp\left(\frac{b}{2x}\left(1 - (1 - 2xt)^{1/2}\right)\right)$$
$$= \sum_{n=0}^{\infty} \left(\frac{b}{2}\right)^n y_n(x; a, b) t^n (n!)^{-1}. \tag{2.4}$$

Their explicit form

$$y_n(x;a,b)=\sum_{k=0}^{n}\binom{n}{k}\binom{n+k+a-2}{k}k!\left(\frac{x}{b}\right)^k \tag{2.5}$$

is given in [12, 19.7, (19)]. The substitution of x, t, a and b, respectively with $2z$, w, 2 and 2 in (2.4) and (2.5), gives the equality

$$\Phi(z,w)=\sum_{n=0}^{\infty}y_n(2z;2,2)w^n(n!)^{-1},$$

i.e.

$$B_n(z)=\frac{1}{n!}\sum_{k=0}^{n}\frac{(n+k)!}{k!(n-k)!}z^k. \tag{2.6}$$

The polynomials $(-1)^n n!B_n(-z)$, which are also called Bessel polynomials, are considered in [13].

Denote

$$\widetilde{B}_n(z)=4^{-n}z^nB_n\left(\frac{1}{z}\right)\exp\left(-\frac{z}{2}\right). \tag{2.7}$$

Having in mind (2.6), we get

$$\widetilde{B}_n(z)=\frac{\exp(-z/2)}{n!4^n}\sum_{k=0}^{n}\frac{(n+k)!}{k!(n-k)!}z^{n-k}. \tag{2.8}$$

Let

$$\widetilde{\Phi}(z,w)=(1-w)^{-1/2}\exp\left\{-\frac{z}{2}(1-w)^{1/2}\right\},\quad z\in\mathbb{C},\ w\in\mathbb{C}\setminus[1,\infty). \tag{2.9}$$

Lemma 2.1. *If* $|w|<1$ *and* $z\in\mathbb{C}$, *then*

$$\widetilde{\Phi}(z,w)=\sum_{n=0}^{\infty}\widetilde{B}_n(z)w^n. \tag{2.10}$$

Proof. The substitutions $z=\zeta^{-1}$ and $w=\zeta\omega/4$ applied consecutively in (2.1), (2.3) give

$$\Phi(\zeta^{-1},w)=(1-4w\zeta^{-1})^{-1/2}\exp\left\{\frac{1-(1-4zw)^{1/2}}{2}\zeta\right\}=\sum_{n=0}^{\infty}B_n(\zeta^{-1})w^n,$$

$$\Phi(\zeta^{-1},\zeta\omega/4)=(1-\omega)^{-1/2}\exp\left\{\frac{1-(1-\omega)^{1/2}}{2}\zeta\right\}=\sum_{n=0}^{\infty}4^{-n}\zeta^nB_n(\zeta^{-1})\omega^n.$$

After multiplication of the last equality by $\exp(-\zeta/2)$ we obtain

$$\exp(-\zeta/2)\Phi(\zeta^{-1},\zeta\omega/4)=(1-\omega)^{-1/2}\exp\left\{-\frac{(1-\omega)^{1/2}}{2}\zeta\right\}=\sum_{n=0}^{\infty}\widetilde{B}_n(\zeta)\omega^n,$$

and since $|4zw|<|\omega|<1$, the lemma is proved.

Denote

$$A_\alpha = \{z : z \in \mathbb{C}^*,\ |\arg z| \le \alpha\pi\}, \quad \mathbb{C}^* = \mathbb{C} \setminus \{0\}. \tag{3.1}$$

Lemma 3.1. *Let $G \subset A_\alpha$, $0 < \alpha < 1$, be a simply connected region, $\gamma \subset G$ be a rectifiable Jordan curve, $F \in H_\gamma$, $F \not\equiv 0$, and $\inf\limits_{z\in\gamma} |z| = r$. Let $|w| < 1/(4r)$ and*

$$f(w) = \int_\gamma F(z)\Phi(z,w)\,dz. \tag{3.2}$$

Then the following expansion holds:

$$f(w) = \sum_{n=0}^{\infty} A_n(F) w^n \tag{3.3}$$

with the coefficients

$$A_n(F) = \int_\gamma F(z) B_n(z)\,dz. \tag{3.4}$$

Moreover, the radius of convergence of the series (3.3) is finite.

Proof. It follows from (2.3) that $B_n(z) = \dfrac{1}{n!}\left\{\dfrac{\partial^n \Phi(z,w)}{\partial w^n}\right\}_{w=0}$. Since $f(w)$ is holomorphic for $|w| < 1/(4r)$, then $f(w)$ can be expanded in a Taylor series

$$f(w) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\int_\gamma F(z)\left\{\frac{\partial^n \Phi(z,w)}{\partial w^n}\right\}_{w=0} dz\right) w^n = \sum_{n=0}^{\infty}\left(\int_\gamma F(z) B_n(z)\,dz\right) w^n,$$

which yield (3.3), if the notations (3.4) are taken into account.

Having in mind the identity (2.2), we get

$$\begin{aligned}\Phi(z,w) &= (1-4zw)^{-1/2} \exp\frac{2w}{1+(1-4zw)^{1/2}} \\ &= (1-4zw)^{-1/2} \exp\left\{-\frac{-2w}{1+(1-4zw)^{1/2}}\right\}.\end{aligned} \tag{3.5}$$

Suppose that the radius of convergence of (3.3) is infinite. This means that (3.3) defines an entire function. Let us evaluate the order of $f(w)$. Using (2.1) and (3.2), we get consecutively

$$\begin{aligned}|f(w)| &\le \int_\gamma \left|F(z)(1-4zw)^{-1/2} \exp\left\{\frac{1-(1-4zw)^{1/2}}{2z}\right\}\right| ds \\ &\le \int_\gamma |F(z)|\,|1-4zw|^{-1/2} \exp\left\{|z|^{-1/2}|w|^{1/2}\left|\frac{w^{-1/2}-(w^{-1}-4z)^{1/2}}{2z^{1/2}}\right|\right\} ds.\end{aligned}$$

As $\lim\limits_{|w|\to\infty}\left|w^{-1/2}-\left(w^{-1}-4z\right)^{1/2}\right|=2|z|^{1/2}$ and $\lim\limits_{|w|\to\infty}(1-4zw)^{-1/2}=0$, then the following inequalities hold:

$$\left|\frac{w^{-1/2}-\left(w^{-1}-4z\right)^{1/2}}{2z^{1/2}}\right|<2,\quad |1-4zw|^{-1/2}<1,$$

for sufficiently large $|w|$. Denoting

$$m=\sup_{z\in\gamma}|F(z)|,\quad \mu(\gamma)=L,\quad M=mL, \tag{3.6}$$

we conclude that there exists a constant $B>0$ such that the inequalities

$$|f(w)|\le M\exp\left(2|z|^{-1/2}|w|^{1/2}\right)\le M\exp\left(r^{-1/2}|w|^{1/2}\right)$$

hold for every $|w|>B$. Therefore, the order of the function f is $\rho\le 1/2$.

Further we apply the Phragmen–Lindelof theorem [14, p. 206] for $f(w)$. To this end, consider first $f(-u)$, $u\ge 0$, and use $\Phi(z,-u)$ as given in (3.5). Since $\gamma\subset A_\alpha$, then $|\arg(1+4zu)|<\alpha\pi$ and $\left|\arg(1+4zu)^{1/2}\right|<\alpha\pi/2$. Therefore $\left|\arg\left(1+(1+4zu)^{1/2}\right)\right|<\alpha\pi/2$. Then $\left|\arg\dfrac{2u}{1+(1+4zu)^{1/2}}\right|<\alpha\pi/2$, i.e. $\operatorname{Re}\dfrac{2u}{1+(1+4zu)^{1/2}}>0$. Further, using the notations $r_1\equiv\inf\limits_{z\in\gamma}\operatorname{Re} z$ and (3.6), we get

$$\begin{aligned}|f(-u)|&\le m\int_\gamma\left|(1+4zu)^{-1/2}\right|\left|\exp\left(-\frac{2u}{1+(1+4zu)^{1/2}}\right)\right|ds\\&\le m(1+4r_1u)^{-1/2}\int_\gamma\exp\left(\operatorname{Re}\frac{-2u}{1+(1+4zu)^{1/2}}\right)ds\\&\le M(1+4r_1u)^{-1/2}.\end{aligned}\tag{3.7}$$

Now, let $\max(\alpha,1-\alpha)<\beta<1$, $\arg(-w)=(1-\beta)\pi$, $\arg z=\theta$. Then $\arg(-zw)=(1-\beta)\pi+\theta$, and as $-\alpha\pi<\theta<\alpha\pi$, we get consecutively

$$\begin{aligned}(1-\alpha-\beta)\pi&<\arg(-zw)<(1+\alpha-\beta)\pi,\\(1-\alpha-\beta)\pi&<\arg(1-4zw)<(1+\alpha-\beta)\pi,\\(1-\alpha-\beta)\frac{\pi}{2}&<\arg(1-4zw)^{1/2}<(1+\alpha-\beta)\frac{\pi}{2}.\end{aligned}$$

Denoting $\psi=\arg\left(1+(1-4zw)^{1/2}\right)$, we have

$$(1-\alpha-\beta)\frac{\pi}{2}<\psi<(1+\alpha-\beta\frac{\pi}{2},\quad \arg\frac{-2w}{1+(1-4zw)^{1/2}}=(1-\beta)\pi-\psi,$$

$$\begin{aligned}(1-\alpha-\beta)\frac{\pi}{2}&=(1-\beta)\pi-(1+\alpha-\beta)\frac{\pi}{2}\\&<(1-\beta)\pi-\psi<(1-\beta)\pi-(1-\alpha-\beta)\frac{\pi}{2}=(1+\alpha-\beta)\frac{\pi}{2},\end{aligned}$$

hence $\left|\arg \dfrac{-2w}{1+(1-4zw)^{1/2}}\right| < \dfrac{\pi}{2}$, i.e. $\operatorname{Re}\left(\dfrac{-2w}{1+(1-4zw)^{1/2}}\right) > 0$. Now, using $\lim\limits_{|w|\to\infty}(1-4zw)^{-1/2} = 0$ and (3.6), we conclude that there exists a constant $P > 0$ such that

$$|f(w)| \le mP \int_\gamma \exp\left\{\operatorname{Re}\left(-\frac{-2w}{1+(1-4zw)^{1/2}}\right)\right\} ds \le MP. \tag{3.8}$$

The rays $l_1 = \{w : w = -u,\ u > 0\}$ and $l_2 = \{w : \arg(-w) = (1-\beta)\pi\}$ divide the complex plane into two angular domains of sizes $(1 \pm \beta)\pi$. The order of the function is $\rho \le 1/2$. It follows from (3.7) and (3.8) that $|f(w)|$ is bounded along l_1 and l_2. As $1/2 < (1 \pm \beta)^{-1}$, according to the Phragmen–Lindelof theorem $f(w)$ is bounded in both angular domains and therefore in the whole complex plane. Hence $f \equiv \text{const}$. It is seen from (3.7) that $\lim\limits_{u\to\infty} f(-u) = 0$, which means $f \equiv 0$. Since $F \not\equiv 0$ and the system $\{B_n(z)\}_{n=0}^{\infty}$ is complete in $H(G)$, see Theorem 1, the last equality contradicts the criterion for completeness (**CC**). Therefore the radius of convergence of the series (3.3) is finite.

Lemma 3.2. *Let $G \subset A_\alpha$, $0 < \alpha < 1/2$, be a simply connected region, $\gamma \subset G$ be a rectifiable curve, $F \in H_\gamma$ and $F \not\equiv 0$. Then there exists a real number $\varphi \in (0, \alpha)$ such that the function f has no singular points outside the set A_φ.*

Proof. The curve γ is a compact set, hence there exists a closed domain A_φ, $0 < \varphi < \alpha$, of the kind (3.1) such that $\gamma \in A_\varphi$ and $\gamma \cap \partial A_\varphi \ne 0$. The values of w, for which $1 - 4zw = 0$, are $w_z = (4z)^{-1}$. Let $z \in \gamma$. Then $w_z \in A_\varphi$ too. Therefore all the points for which $1 - 4zw = 0$ are in the set A_φ and the function $(1-4zw)^{-1/2}$ is a holomorphic function of w outside A_φ. Hence the function (3.2) is holomorphic for $w \in \operatorname{Ext} A_\varphi$ too.

Lemma 3.3. *Let $G \subset A_\alpha$, $0 < \alpha < 1/2$, be a simply connected region, $\gamma \subset G$ be a rectifiable Jordan curve, $F \in H_\gamma$ and $F \not\equiv 0$. Let*

$$\widetilde{f}(w) = \int_\gamma F(z)\widetilde{\Phi}(z, w)\, dz, \quad w \in \mathbb{C} \setminus [1, \infty). \tag{3.9}$$

Then the following expansion holds:

$$\widetilde{f}(w) = \sum_{n=0}^{\infty} \widetilde{A}_n(F) w^n \tag{3.10}$$

for $|w| < 1$ with coefficients

$$\widetilde{A}_n(F) = \int_\gamma F(z)\widetilde{B}_n(z)\, dz. \tag{3.11}$$

Moreover, the radius of convergence of the series of (3.10) is finite and it has no singular points in $\mathbb{C} \setminus [1, \infty)$.

Proof. From (2.10) it follows that $\widetilde{B}_n(z) = \dfrac{1}{n!}\left\{\dfrac{\partial^n \widetilde{\Phi}(z,w)}{\partial w^n}\right\}_{w=0}$. As $\widetilde{f}(w)$ is holomorphic for $|w| < 1$, then $\widetilde{f}(w)$ can be expanded in a Taylor series, i.e.:

$$\widetilde{f}(w) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\int_\gamma F(z) \left\{\frac{\partial^n \widetilde{\Phi}(z,w)}{\partial w^n}\right\}_{n=0} dz\right) w^n = \sum_{n=0}^{\infty}\left(\int_\gamma F(z)\widetilde{B}_n(z)\,dz\right) w^n,$$

which yields (3.10), if the notations (3.11) are taken into account.

Suppose that (3.10) has infinite radius of convergence. This means that (3.10) defines the entire function $\widetilde{f}$. From (2.9) and (3.9) we obtain

$$\left|\widetilde{f}(w)\right| \le \int_\gamma |F(z)|\,|1-w|^{-1/2} \exp\left\{\frac{|z|}{2}\,|w|^{1/2}\left|w^{-1}-1\right|^{1/2}\right\} ds.$$

Since $\lim\limits_{|w|\to\infty} \left|w^{-1}-1\right|^{1/2} = 1$ and $\lim\limits_{|w|\to\infty} |1-w|^{-1/2} = 0$, then the inequalities $\left|w^{-1}-1\right|^{1/2} < 2$, $|1-w|^{-1/2} < 1$ hold for sufficiently large $|w|$. If we denote $R = \sup\limits_{z\in\gamma} |z|$ and use (3.6), we obtain that there exists a constant $D > 0$ such that the inequality $\left|\widetilde{f}(w)\right| \le M \exp\left(R|w|^{1/2}\right)$ holds for every $|w| > D$. This means that $\widetilde{f}$ is of order $\rho \le 1/2$.

Now let us investigate the behaviour of $\widetilde{f}(w)$ along each of the rays $l_1 = \{w : w = -u,\ u > 0\}$ and $l_3 = \{w : \arg(-w) = (1-2\alpha)\pi/2\}$. As $\gamma \subset A_\alpha$, then $\left|\arg\left(\dfrac{z}{2}(1+u)^{1/2}\right)\right| < \alpha\pi$, i.e. $\operatorname{Re}\left(\dfrac{z}{2}(1+u)^{1/2}\right) > 0$. Using the notation (3.6), we get

$$\begin{aligned}\left|\widetilde{f}(-u)\right| &\le m(1+u)^{-1/2} \int_\gamma \exp\left\{\operatorname{Re}\left(-\frac{z}{2}(1+u)^{1/2}\right)\right\} ds \\ &\le M(1+u)^{-1/2} \le M. \end{aligned} \tag{3.12}$$

Now let $w \in l_3$. As $-\alpha\pi < \arg z < \alpha\pi$, we have consecutively

$$0 < \arg(1-w) < (1-2\alpha)\frac{\pi}{2},$$

$$0 < \arg(1-w)^{1/2} < (1-2\alpha)\frac{\pi}{4},$$

$$0 < \arg\left(\frac{z}{2}(1-w)^{1/2}\right) < (1+2\alpha)\frac{\pi}{4}, \quad \text{i.e. } \operatorname{Re}\left(\frac{z}{2}(1-w)^{1/2}\right) > 0.$$

Using that $\lim\limits_{|w|\to\infty} |1-w|^{-1/2} = 0$ and (3.6), we conclude that there exists a constant $Q > 0$ such that

$$\left|\widetilde{f}(w)\right| \le mQ \int_\gamma \exp\left\{\operatorname{Re}\left(-\frac{z}{2}(1-w)^{1/2}\right)\right\} ds \le MQ. \tag{3.13}$$

The rays l_1 and l_3 divide the complex plane into two angular domains with sizes $(1-2\alpha)\pi/2$ and $(3+2\alpha)\pi/2$. The order of the function is $\rho \leq 1/2$. As seen from (3.12) and (3.13), $\widetilde{f}(w)$ is bounded along l_1 and l_3. Because of $1/2 < 2(1-2\alpha)^{-1}$ and $1/2 < 2(3+2\alpha)^{-1}$, according to the Phragmen-Lindelof theorem $\widetilde{f}(w)$ is bounded in both angular domains and therefore on the whole complex plane. Hence $\widetilde{f} \equiv \text{const}$. From (3.12) it is seen that $\lim\limits_{u\to\infty} \widetilde{f}(-u) = 0$, that is $\widetilde{f} \equiv 0$. Since $F \not\equiv 0$ and the system $\left\{\widetilde{B}_n(z)\right\}_{n=0}^{\infty}$ is complete in $H(G)$, see Theorem 2, the last equality contradicts the criterion **(CC)**. Therefore the series (3.10) has a finite radius of convergence. Finally, let us note that (3.9) has no singular points in $\mathbb{C} \setminus [1,\infty)$.

4. MAIN RESULTS

Theorem 4.1. *Let $G \subset \mathbb{C}$ be a simply connected region. Then:*

i) *The system of the polynomials $\{B_n(z)\}_{n=0}^{\infty}$ is complete in the space $H(G)$;*

ii) *The system of the functions $\left\{\widetilde{B}_n(z)\right\}_{n=0}^{\infty}$ is complete in the space $H(G)$.*

Proof. i) According to (2.6) $\deg B_n = n$, $n = 0,1,2,\ldots$, and therefore the system $\{B_n(z)\}_{n=0}^{\infty}$ is linearly independent. Therefore $\{B_n(z)\}_{n=0}^{\infty}$ is a basis in the space of the algebraic polynomials. Hence z^n is a linear combination of $\{B_k(z)\}_{k=0}^{n}$, therefore it can be concluded that $\{B_n(z)\}_{n=0}^{\infty}$ is complete in $H(G)$.

ii) According to (2.8) the coefficients of the polynomials $\exp(z/2)\widetilde{B}_n(z)$ are all different from zero, i.e. $\deg\left(\exp(z/2)\widetilde{B}_n(z)\right) = n$, $n = 0,1,2,\ldots$ Therefore the system $\left\{\exp(z/2)\widetilde{B}_n(z)\right\}_{n=0}^{\infty}$ is linearly independent, which means that it is a basis in the space of algebraic polynomials. Then z^n is a linear combination of $\left\{\exp(z/2)\widetilde{B}_k(z)\right\}_{k=0}^{n}$. That is why $\left\{\exp(z/2)\widetilde{B}_n(z)\right\}_{n=0}^{\infty}$ is compete in $H(G)$, and since $\exp(z/2) \neq 0$ for each $z \in \mathbb{C}$, the correctness of the theorem is proved.

Theorem 4.2. *Let $0 < \alpha < 1$ and $\lim\limits_{n\to\infty}(n/k_n) = \delta \geq \alpha$. Then the system of the polynomials*

$$\{B_{k_n}(z)\}_{n=0}^{\infty} \tag{4.1}$$

is complete in the space $H(G)$ for each simply connected region $G \subset A_\alpha$.

Proof. Suppose the statement is not correct. Then there exists a simply connected region $G \subset A_\alpha$ such that the system (4.1) is not complete in $H(G)$. According to the criterion **(CC)** this means that there exist a rectifiable Jordan curve $\gamma \subset G$ and a function $G \in H_\gamma$ such that $F \not\equiv 0$, but

$$\int_\gamma F(z)B_{k_n}(z)\,dz = 0, \quad n = 0,1,2,\ldots \tag{4.2}$$

Let $r = \inf\limits_{z \in \gamma} |z|$ and $|w| < (4r)^{-1}$. Consider the complex-valued function $f(w)$, defined in (3.2). Let us note that it is not identically zero. Moreover, if $\widetilde{k}_n$ are the indices of the coefficients (3.4) in the power series (3.3), for which $\left\{\widetilde{k}_n\right\}_{n=0}^{\infty} = \{n\}_{n=0}^{\infty} \setminus \{k_n\}_{n=0}^{\infty}$, it follows from (4.2) that

$$f(w) = \sum_{n=0}^{\infty} A_{\widetilde{k}_n}(F) w^{\widetilde{k}_n}. \tag{4.3}$$

For the density of the sequence $\left\{\widetilde{k}_n\right\}_{n=0}^{\infty}$ we have

$$\Delta = 1 - \delta \leq 1 - \alpha. \tag{4.4}$$

As $F \not\equiv 0$, not all the complex numbers (3.4) are zeroes. Then, according to Lemma 2, there exists a number $\varphi \in (0, \alpha)$ such that all singular points on the circle $|w| = R$ (R is the radius of the convergence of the series (3.3)) lie in the set A_φ, i.e. there is a closed arc with lenght $2\pi(1-\varphi)$, where (3.3) has no singular points. On the other hand, by a Polya theorem [15, Th. 7, p. 625] every closed arc of the circle $|w| = R$ with lenght $2\pi\Delta$ contains at least one singular point of (4.3). Because of (4.4) we have $2\pi\Delta = 2\pi(1-\delta) \leq 2\pi(1-\alpha) < 2\pi(1-\varphi)$ and we come to a contradiction. Therefore the system (4.1) is complete in $H(G)$ for every simply connected region $G \subset A_\alpha$.

Theorem 4.3. *Let* $0 < \alpha < 1/2$ *and* $\lim\limits_{n \to \infty} (n/k_n) = \delta > 0$. *Then the system of the functions*

$$\left\{\widetilde{B}_{k_n}(z)\right\}_{n=0}^{\infty} \tag{4.5}$$

is complete in the space $H(G)$ *for every simply connected region* $G \subset A_\alpha$.

Proof. Let us suppose that the statement is not correct. Then there exists a simply connected region $G \subset A_\alpha$ such that the system (4.5) is not complete in $H(G)$. That means that there exist a rectifiable Jordan curve $\gamma \subset G$ and a function $F \in H_\gamma$ such that $F \not\equiv 0$, but

$$\int_\gamma F(z) \widetilde{B}_{k_n}(z)\, dz = 0, \quad n = 0, 1, 2, \ldots \tag{4.6}$$

Let $|w| < 1$. Consider the complex-valued function $\widetilde{f}(w)$, defined by the equality (3.9). Observe that it is not identically zero. Moreover, if $\widetilde{k}_n$ are the indices of the coefficients (3.11) in the power series (3.10) for which $\left\{\widetilde{k}_n\right\}_{n=0}^{\infty} = \{n\}_{n=0}^{\infty} \setminus \{k_n\}_{n=0}^{\infty}$, it follows from (4.6) that

$$\widetilde{f}(w) = \sum_{n=0}^{\infty} \widetilde{A}_{\widetilde{k}_n}(F) w^{\widetilde{k}_n}. \tag{4.7}$$

We have

$$\Delta = 1 - \delta < 1 \tag{4.8}$$

for the density of the sequence $\left\{\widetilde{k}_n\right\}_{n=0}^{\infty}$. As $F \not\equiv 0$, not all of the complex numbers (3.11) are equal to zero. Then, according to Lemma 3, the unique singular point of $f(w)$ on the circle $|w| = R$ (R is the radius of convergence of the series (3.10)) is $w = R$. On the other hand, according to a Polya theorem [15], every closed arc of the circle $|w| = R$ with lenght $2\pi\Delta$ contains at least one singular point of (4.7). Because of (4.8) we have $2\pi\Delta = 2\pi(1-\delta) < 2\pi$ and we come to a contradiction. Therefore the system (4.5) is complete in $H(G)$ for every simply connected region $G \subset A_\alpha$.

Acknowledgements. The author is thankful to Prof. P. Rusev for the interest shown in these results and the useful recommendations.

REFERENCES

1. Zygmund, A., S. Saks. Analytic Functions. Warszawa-Wroclaw, 1952.
2. Levin, B. Distribution of Zeros of Entire Functions. Providence, 1964.
3. Казьмин, Ю. О подпоследовательностях полиномов Эрмита и Лагерра. — Вестн. Моск. унив. **2**, 1960, 6–9.
4. Леонтьев, А. Последовательности полиномов из экспонент. Москва, 1980.
5. Русев, П. О полноте системы функций Лагерра второго рода. — Докл. БАН, **30**, N1, 1977, 9–11.
6. Rusev, P. Completeness of Laguerre and Hermite Functions of Second Kind. Constructive Function Theory'77, Sofia, 1980, 469–473.
7. Rusev, P. Complete Systems of Jacobi Associated Functions in Spaces of Holomorphic Functions. — Analysis, **14**, Munchen, 1994, 249–255.
8. Rusev, P. Complete Systems of Tricomi Functions in Spaces of Holomorphic Functions. — Год. Соф. унив., ФМИ, **88**, кн. 1, 1994.
9. Rusev, P. Complete Systems of Kummer and Weber–Hermite Functions in Spases of Holomorphic Functions. In: Symposia Gaussiana, Conf. Berlin-New York, 1995, 723–731.
10. Boas, P., R. Buck. Polynomial Expansion of Analytic Function. Berlin-Göttingen-Heidelberg, 1958.
11. Burchnall, J. The Bessel Polynomials. — Canad. J. Math., **3**, 1951, 62–68.
12. Erdelyi, A., et al. Higher Trancendental Functions. McGraw-Hill, New York-Toronto-London, 1953.
13. Обрешков, Н. Върху нлкои ортогонални полиноми в комплексна област. — Изв. Мат. инст., **2**, кн. 1, 1956, 45–67.
14. Маркушевич, А. Теория аналитических функций. Т. 2, М., 1968.
15. Polya, G. Unterchungen uber Lucken und Singularitaten von Potenzreihen. — Math. Zeitschr., 29, N4, 1929, 549–640.

Received on 19.07.1996

Institute of Applied Mathematics and Informatics
Technical University
1156 Sofia
Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

AN ALGORITHMIC APPROACH TO SOME PROBLEMS ON THE REPRESENTATION OF NATURAL NUMBERS AS SUMS WITHOUT REPETITIONS[1]

DIMITER SKORDEV

Given any strictly increasing computable function in the set of natural numbers, certain algorithmic problems arise on the representation of numbers as sums of distinct values of the function. The problem whether a given natural number is representable in this form is obviously algorithmically solvable, but we propose some methods for the solution of the problem that seem to be better than the straightforward ones.

It is easy to see the algorithmic unsolvability of the problem whether all natural numbers are representable (under the usual assumption that an index of the given computable function is used as input data). However, under an appropriate restriction concerning, roughly speaking, the speed of the growth of the function, we present an algorithm for solving this problem and the more general one whether all natural numbers greater than a given one are representable (the restriction is satisfied, for example, when the given function is a polynomial).

We make applications of the presented positive results to concrete problems concerning, for instance, the representation as sums of distinct squares or as sums of distinct positive cubes.

Keywords: algorithm, sums without repetitions, representability of natural numbers, sums of distinct squares, sums of distinct positive cubes.
Mathematics Subject Classification: 11–04, 11B13, 11E25.

1. INTRODUCTION

Let $\mathbf{N}_+$ be the set of the positive integers. Suppose f is a strictly increasing function in $\mathbf{N}_+$. An integer n will be called *additively f–representable without*

[1] Lecture presented at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

repetitions (*f–representable*, for short) iff

$$n = \sum_{i \in A} f(i)$$

for some finite subset A of $\mathbf{N}_+$; any such A will be called an *f–representation of n*. Of course, all f–representable integers are non-negative, and the number 0 is f–representable (with an empty f–representation).

There is a case when any non-negative integer is f–representable and has a unique f–representation. This is the case when $f(i) = 2^{i-1}$ for $i = 1, 2, 3, \ldots$ To have a more complicated example concerning f–representability, let us consider the case when $f(i) = i^2$ for any i in $\mathbf{N}_+$. Then there exist positive integers that are not f–representable, as well as ones having more than one f–representation. Some results connected to f–representability in this case have been presented in [2–5], but without giving a complete description of the set of the representable integers. Such a description can be derived from certain results given in [1] that show the f–representability of all integers greater than 128 as well as of the most of the smaller positive integers. By checking individually the few remaining positive integers, one gets the following conclusion: there are exactly 31 positive integers that are not f–representable, namely the integers 2, 3, 6, 7, 8, 11, 12, 15, 18, 19, 22, 23, 24, 27, 28, 31, 32, 33, 43, 44, 47, 48, 60, 67, 72, 76, 92, 96, 108, 112, 128.

The mentioned results from [1] are proved by using tools from Number Theory (such as, e.g., divisibility considerations). Those results give in fact considerably more precise information about the f–representations in question. For example, it is seen that each f–representable integer in the considered case has an f–representation consisting of not more than six elements. However, it could be possibly interesting to know that the less precise statement, formulated at the end of the previous paragraph, can be proved in an algorithmic way without using any specific tools from Number Theory. This can be done as an application of a certain method that will be exposed in the present paper.

2. A USEFUL EXTENSION OF THE INVERSE FUNCTION f^{-1}

We turn back to the general case described in the first paragraph of the introduction. Given the function f, we define three other functions REPR_f, L_f and $f^\dagger$ with domain $\mathbf{N}_+$, the first two of them being set-valued and the third one being integer-valued. We define them as follows. Let n be an arbitrary element of $\mathbf{N}_+$. We adopt $\mathsf{REPR}_f(n)$ to be the set of all f–representations of n; clearly, this set is finite (possibly empty) and its elements (if any) are non-empty finite subsets of $\mathbf{N}_+$. Then we set

$$L_f(n) = \{\min A \mid A \in \mathsf{REPR}_f(n)\}.$$

Of course, $L_f(n)$ is a finite subset of $\mathbf{N}_+$, and $L_f(n)$ is empty iff $\mathsf{REPR}_f(n)$ is empty, i.e. iff n is not f–representable. Finally, if $L_f(n) \neq \emptyset$, then we set $f^\dagger(n)$ to be the maximal element of $L_f(n)$, otherwise we set $f^\dagger(n) = 0$. Thus $f^\dagger(n)$ is a non-negative integer that is equal to 0 iff n is not f–representable.

Example 1. If $f(i) = i^2$ for $i = 1, 2, 3, \ldots$, then

$$\mathsf{REPR}_f(50) = \{\{1,7\},\{1,2,3,6\},\{3,4,5\}\},$$

hence $L_f(50) = \{1,3\}$, $f^\dagger(50) = 3$.

For any positive integer i the singleton $\{i\}$ is an f–representation of the number $f(i)$, and any f–representation of this number contains some element not greater than i, hence the equality $f^\dagger(f(i)) = i$ holds. Thus the function $f^\dagger$ is an extension of the inverse function f^{-1}.

We also note that for any f–representable positive integer n the number $f^\dagger(n)$ belongs to some f–representation of n, hence the inequality $f(f^\dagger(n)) \le n$ holds.

The consecutive values of the function $f^\dagger$ can be recursively computed on the base of the next proposition.

Theorem 1. *For any positive integer n we have the equality*

$$L_f(n) = \{k \in \mathbf{N}_+ \mid f(k) = n \text{ or } (f(k) < n \text{ and } f^\dagger(n - f(k)) > k)\}.$$

Proof. Let n be a positive integer. Consider first any k belonging to $L_f(n)$. Then $k = \min A$ for some f–representation A of n, hence $k \in \mathbf{N}_+$. If k is the only element of A, then $f(k) = n$. Otherwise $n - f(k)$ is a positive integer, and $A \setminus \{k\}$ is an f-representation of $n - f(k)$. Therefore

$$f^\dagger(n - f(k)) \ge \min(A \setminus \{k\}) > k.$$

Thus in both cases k belongs to the right-hand side of the equality. For the reasoning in the opposite direction, suppose now that k belongs to the right-hand side of this equality. Then $k \in \mathbf{N}_+$, and either $f(k) = n$ or $f(k) < n$ and $f^\dagger(n - f(k)) > k$. If $f(k) = n$, then we set $A = \{k\}$. Otherwise we consider an f-representation B of $n - f(k)$ such that $f^\dagger(n - f(k)) = \min B$, and we set $A = \{k\} \cup B$. In both cases A is an f–representation of n and $k = \min A$, hence $k \in L_f(n)$.

Example 2. Let f enumerate the set of the prime numbers, i.e. $f(1) = 2$, $f(2) = 3$, $f(3) = 5$, $f(4) = 7$, $f(5) = 11$ and so on. Then, making use of Theorem 1 and of the definition of the function $f^\dagger$, we get consecutively:

$$\begin{aligned}
L_f(1) &= \emptyset, & f^\dagger(1) &= 0,\\
L_f(2) &= \{1\}, & f^\dagger(2) &= 1,\\
L_f(3) &= \{2\}, & f^\dagger(3) &= 2,\\
L_f(4) &= \emptyset, & f^\dagger(4) &= 0,\\
L_f(5) &= \{1,3\}, & f^\dagger(5) &= 3,\\
L_f(6) &= \emptyset, & f^\dagger(6) &= 0,\\
L_f(7) &= \{1,4\}, & f^\dagger(7) &= 4,\\
L_f(8) &= \{2\}, & f^\dagger(8) &= 2,\\
L_f(9) &= \{1\}, & f^\dagger(9) &= 1,\\
L_f(10) &= \{1,2\}, & f^\dagger(10) &= 2.
\end{aligned}$$

x	$f^\dagger(10x+y)$									
	$y=0$	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$	$y=6$	$y=7$	$y=8$	$y=9$
0		1	0	0	2	1	0	0	0	3
1	1	0	0	2	1	0	4	1	0	0
2	2	1	0	0	0	5	1	0	0	2
3	1	0	0	0	3	1	6	1	2	1
4	2	4	1	0	0	3	1	0	0	7
5	3	1	4	2	2	1	2	1	3	1
6	0	5	2	1	8	4	1	0	2	2
7	3	1	0	3	5	1	0	4	2	1
8	4	9	1	3	2	6	3	2	1	5
9	4	1	0	2	3	1	0	4	3	3
10	10	4	2	2	2	4	5	1	0	3
11	5	1	0	7	3	3	4	6	2	3
12	2	11	4	2	1	5	4	1	0	4

Fig. 1. The first 129 values of the function $f^\dagger$ in the case of $f(i)=i^2$

Clearly, it is not always necessary to find all elements of the set $L_f(n)$ in order to see that it is not empty and to find its maximal element. We have $f(k) \leq n$ for any k in $L_f(n)$. Therefore, to calculate $f^\dagger(n)$, one could simply find the least positive integer k such that $f(k) > n$ and then execute the operator

$$\textbf{repeat } k := k-1 \textbf{ until } k = 0 \text{ or } k \in L_f(n)$$

(interpreted in a Pascal-like way).

Example 3. Fig. 1 contains a table of the values of $f^\dagger(n)$ for $n = 1, 2, \ldots, 129$, calculated by computer in the above way in the case of $f(i) = i^2$. The table shows that among the positive integers not greater than 129, exactly the 31 ones listed in the introduction are not f–representable.

The amount of operations can be somewhat reduced by noticing that for positive integers n, not belonging to the range of f, one could start executing the above operator from the least positive integer k such that $f(k) \geq n/2$ (if $n \in \mathbf{N}_+ \setminus \text{range}(f)$, then $f(k) < n/2$ for any k in $L_f(n)$, since any k in $L_f(n)$ belongs to some f–representation of n together with at least one greater number). Working in this way, one could manually verify the correctness of the values in Fig. 1 in the course of, let us say, one and a half hour.

Let $\mathbf{N}$ be the set of all non-negative integers. The indicated method for computing values of the function $f^\dagger$ can be modified by introducing a binary relation H_f in $\mathbf{N}$ as follows: $n\ H_f\ i$ iff n has an f–representation A such that all elements of A are greater than i. We have $0\ H_f\ i$ for any i in $\mathbf{N}$ by trivial reasons. On the other hand, the following equivalence holds for any n in $\mathbf{N}_+$ and any i in $\mathbf{N}$: $n\ H_f\ i$ iff $f^\dagger(n) > i$. Making use of these properties of H_f and of Theorem 1, we get the following result.

Theorem 2. *Let* $n \in \mathbf{N}_+$. *Then*

$$L_f(n) = \{k \in \mathbf{N}_+ \mid f(k) \leq n \text{ and } n - f(k)\ H_f\ k)\}$$

and for any i *in* $\mathbf{N}$

$$n\ H_f\ i \Leftrightarrow \exists k \in \mathbf{N}_+\ (k > i \text{ and } f(k) \leq n \text{ and } n - f(k)\ H_f\ k).$$

To illustrate the application of the relation H_f to the computation of values of $f^\dagger$, we shall consider one more example.

Example 4. Let, as in Examples 1 and 3, $f(i) = i^2$ for $i = 1, 2, 3, \ldots$ We shall compute $f^\dagger(50)$ by using the properties of the relation H_f. Since 50 is not a value of the function f and the least positive integer k satisfying the inequality $k^2 \geq 50/2$ is 5, the value of $f^\dagger(50)$ can be obtained from $k = 5$ by applying the operator

$$\textbf{repeat } k := k - 1 \textbf{ until } k = 0 \text{ or } k \in L_f(50).$$

By Theorem 2 we have

$$4 \in L_f(50) \Leftrightarrow 50 - 4^2\ H_f\ 4 \Leftrightarrow 34\ H_f\ 4 \Leftrightarrow$$

$$\exists k \in \mathbf{N}_+\ (k > 4 \text{ and } k^2 \leq 34 \text{ and } 34 - k^2\ H_f\ k) \Leftrightarrow 34 - 5^2\ H_f\ 5 \Leftrightarrow$$

$$9\ H_f\ 5 \Leftrightarrow \exists k \in \mathbf{N}_+\ (k > 5 \text{ and } k^2 \leq 9 \text{ and } 9 - k^2\ H_f\ k),$$

hence $4 \notin L_f(50)$. Again by Theorem 2

$$3 \in L_f(50) \Leftrightarrow 50 - 3^2\ H_f\ 3 \Leftrightarrow 41\ H_f\ 3 \Leftrightarrow$$

$$\exists k \in \mathbf{N}_+\ (k > 3 \text{ and } k^2 \leq 41 \text{ and } 41 - k^2\ H_f\ k) \Leftrightarrow$$

$$41 - 4^2\ H_f\ 4 \text{ or } 41 - 5^2\ H_f\ 5 \text{ or } 41 - 6^2\ H_f\ 6,$$

$$41 - 4^2\ H_f\ 4 \Leftrightarrow 25\ H_f\ 4 \Leftrightarrow$$

$$\exists k \in \mathbf{N}_+\ (k > 4 \text{ and } k^2 \leq 25 \text{ and } 25 - k^2\ H_f\ k) \Leftrightarrow$$

$$25 - 5^2\ H_f\ 5 \Leftrightarrow 0\ H_f\ 5,$$

hence $41 - 4^2\ H_f\ 4$, and therefore $3 \in L_f(50)$. Thus $f^\dagger(50) = 3$.

Remark. The method used in the above example is convenient when some value of the function $f^\dagger$ has to be computed without necessarily computing the preceding ones (an additional reduction of the count of the operations could be achieved by noticing that the statements of Theorem 2, in particular the second one, hold also with "$f(k) < n/2$" instead of "$f(k) \leq n$" in the case of $n \in \mathbf{N}_+ \setminus \text{range}(f)$). However, if one has to make a table of the values of $f^\dagger(n)$ for $n = 1, 2, \ldots, m$, where m is a given positive integer, then it seems more reasonable to proceed by consecutive straightforward applications of Theorem 1 as in Example 3.

The function $f^\dagger$ can be used not only for checking whether a given positive integer is f–representable, but also for finding one of the f–representations of a given f-representable natural number. This way of using $f^\dagger$ is possible on the basis of the next proposition.

Theorem 3. *Let n be an f–representable non-negative integer. Let the integers $n_0, n_1, \ldots$ be defined as follows, taken for granted that n_{j+1} is defined iff the right-hand side of the second equality makes sense:*

$$n_0 = n, \quad n_{j+1} = n_j - f(f^\dagger(n_j)).$$

Then there is a non-negative integer r such that $n_r = 0$, and if r is such an integer, then the set $\{f^\dagger(n_j) \mid 0 \leq j < r\}$ is an f-representation of n.

Proof. It is clear that n_{j+1} is defined iff n_j is positive and f–representable. Hence, if n_r is defined for a certain r, then n_j is defined, positive and f–representable for any $j < r$, and if $n_r = 0$, then n_j is undefined for any $j > r$. Applying the last statement in the case of $r = 0$, we see that the theorem is trivial if $n = 0$. Suppose now that $n > 0$. Then n_0 is positive and f–representable. On the other hand, if for a certain j the number n_j is defined, positive and f–representable, then, by the definition of the function $f^\dagger$, the number n_{j+1} is not only defined, but it has an f–representation whose elements are all greater than $f^\dagger(n_j)$, and this implies the inequality $f^\dagger(n_j) < f^\dagger(n_{j+1})$ in the case of $n_{j+1} > 0$. Since the values of the function f are positive, we thus see that the defined numbers n_j form a strictly decreasing sequence of f–representable and hence non-negative integers, and the defined numbers $f^\dagger(n_j)$ form a strictly increasing sequence. The sequence $n_0, n_1, \ldots$ should be necessarily finite, and it is clear that its last member should be 0. Consider now an r such that $n_r = 0$, and set $A = \{f^\dagger(n_j) \mid 0 \leq j < r\}$. Then

$$n = n_0 - n_r = \sum_{j=0}^{r-1}(n_j - n_{j+1}) = \sum_{j=0}^{r-1} f(f^\dagger(n_j)) = \sum_{i \in A} f(i).$$

Hence A is an f–representation of n.

Example 5. We shall apply the above theorem to $f(i) = i^2$ and $n = 124$. In this case we get (using the table from Fig. 1)

$$n_0 = 124,\ f^\dagger(n_0) = 1,\ n_1 = 123,\ f^\dagger(n_1) = 2,\ n_2 = 119,\ f^\dagger(n_2) = 3,$$

$$n_3 = 110,\ f^\dagger(n_3) = 5,\ n_4 = 85,\ f^\dagger(n_4) = 6,\ n_5 = 49,\ f^\dagger(n_5) = 7,\ n_6 = 0.$$

Hence, by Theorem 3, the set $\{1, 2, 3, 5, 6, 7\}$ is an f–representation of the number 124.

3. CHECKING IF ALL NATURAL NUMBERS GREATER THAN A GIVEN ONE ARE f-REPRESENTABLE

As until now, a strictly increasing function f from $\mathbf{N}_+$ to $\mathbf{N}_+$ is supposed to be given. If this function is computable (in the precise sense given by Recursive Function Theory), then there are obvious algorithms solving the problem whether a given natural number is f-representable, and the considerations from the previous section yield certain better algorithms for the same purpose. A more difficult problem is to decide whether all natural numbers are f-representable. This problem is algorithmically unsolvable in the following natural sense: there is no computable function h defined on the indices of all strictly increasing computable functions f in $\mathbf{N}_+$ and transforming such an index into 0 exactly when all natural numbers are f-representable with respect to the corresponding function f. To prove this, let us consider a two-argument primitive recursive function g such that the set $P = \{x \mid \exists y\,(g(x,y) = 0)\}$ is not recursive. For each x in $\mathbf{N}$ we define a strictly increasing function f_x from $\mathbf{N}_+$ into $\mathbf{N}_+$ as follows: for any i in $\mathbf{N}_+$, if $g(x,y) > 0$ for all y less than i, then $f_x(i) = 2^{i-1}$, otherwise $f_x(i) = 2^i$. If $x \in P$, then the range of the corresponding function f_x is the set $\{1, 2, 2^2, 2^3, \ldots\}$ with one of its elements missing, otherwise the range of f_x is the whole set $\{1, 2, 2^2, 2^3, \ldots\}$. Hence, if $x \in P$, then there are infinitely many natural numbers that are not f_x-representable, otherwise all natural numbers are f_x-representable. If we suppose that a computable function h exists telling apart indices as said above, then we get a contradiction with the non-recursiveness of P.

Of course, the established algorithmic unsolvability directly implies the unsolvability of the more general problem to decide whether all natural numbers greater than a given one are f-representable. However, we cannot exclude the possibility of an algorithmic solution of the last problem under some reasonable restrictions imposed on the function f. A realization of this possibility will be demonstrated in the present section.

For any two integers a and b let $[a\,..\,b)$ denote the set of all integers n satisfying the inequalities $a \leq n < b$ (of course, this set is non-empty iff $a < b$). Let $[a\,..\,\infty)$ denote the set of all integers n satisfying the inequality $a \leq n$.

Theorem 4. *Suppose $i_0 \in \mathbf{N}_+$, $n_0 \in \mathbf{N}$, and the following two conditions are satisfied:*

1. *For any i in $[i_0\,..\,\infty)$ the inequality $2f(i) - f(i+1) \geq n_0$ holds.*
2. *All elements of $[n_0\,..\,n_0 + f(i_0))$ are f-representable.*

Then all elements of $[n_0\,..\,\infty)$ are f-representable.

Proof (making use of an idea from [5]). For any positive integer i we set $S_i = [n_0 + f(i)\,..\,2f(i))$. We shall first show that any element of the set $[n_0 + f(i_0)\,..\,\infty)$ belongs to some S_i (with $i \geq i_0$). In fact, given an element n of $[n_0 + f(i_0)\,..\,\infty)$, let us consider the greatest i in $\mathbf{N}_+$ satisfying the inequality $n_0 + f(i) \leq n$. For that i we have the inequalities $i \geq i_0$, $n_0 + f(i+1) > n$. From them and Condition 1, the inequality $n < 2f(i)$ follows, hence $n \in S_i$. Now we shall prove the conclusion of the theorem by means of an induction of the following kind: we shall show that

whenever an integer n belongs to the set $[n_0 .. \infty)$ and all smaller integers belonging to this set are f–representable, then n is also f–representable. Suppose n is an integer satisfying the above assumptions; we shall prove that n is f–representable. By Condition 2, we have to examine only the case when $n \geq n_0 + f(i_0)$. Then we consider a positive integer i such that $n \in S_i$. The last condition is equivalent to the inequalities $n_0 \leq n - f(i) < f(i)$. The first of them, together with the inequality $n - f(i) < n$ and the induction hypothesis, shows that $n - f(i)$ is f–representable. Let A be an f–representation of $n - f(i)$. The inequality $n - f(i) < f(i)$ implies that $i \notin A$. This fact, combined with the equality $n = (n - f(i)) + f(i)$, shows that $A \cup \{i\}$ is an f–representation of n, hence n is f–representable.

Remark. An inspection of the proof shows that Condition 2 may be weakened by requiring f–representability only of the elements of $[n_0 .. n_0 + f(i_0))$ that belong to none of the sets S_i, $i = 1, 2, 3, \ldots$

Suppose now some $n_0 \in \mathbf{N}$ is given. Theorem 4 immediately implies the following statement: whenever $i_0 \in \mathbf{N}_+$ and Condition 1 is satisfied, then the f–representability of all elements of $[n_0 .. \infty)$ is equivalent to the representability of the elements of $[n_0 .. n_0 + f(i_0))$. If the function f is computable, then the last condition can be checked in an algorithmic way, and this will be an algorithmic way to check whether all elements of $[n_0 .. \infty)$ are f–representable. Of course, we may use this way only if we succeed to find some $i_0 \in \mathbf{N}_+$ satisfying Condition 1. We shall show now some examples when such an i_0 really can be found.

Example 6. Let $f(i) = 2^{i-1}$ for $i = 1, 2, 3, \ldots$ Then $2f(i) - f(i+1) = 0$ for any such i, hence Condition 1 is satisfied with $n_0 = 0$, $i_0 = 1$. Therefore the well-known f–representability of all non-negative integers in this case can be proved by checking the f–representability of the elements of the set $[0 .. f(1))$. Thus the f–representability of all non-negative integers is reduced to the trivial fact that 0 is f–representable.

Example 7 (generalization of the previous example). If $2f(i) - f(i+1) \geq 0$ for any i, then the f–representability of all non-negative integers is equivalent to the equality $f(1) = 1$ (since no f–representable positive integer can be less than $f(1)$). As a particular instance of this we could consider the case when f enumerates the Fibonacci numbers $1, 2, 3, 5, 8, 13, \ldots$, i.e. $f(1) = 1$, $f(2) = 2$ and $f(i) = f(i-1) + f(i-2)$ for $i = 3, 4, 5, \ldots$ In this case, if $i = 1$, then $2f(i) - f(i+1) = 0$, otherwise $2f(i) - f(i+1) = f(i) - f(i-1) > 0$. Thus all non-negative integers are f-representable with respect to this particular function f.

Example 8. Let the function f be a polynomial, i.e.

$$f(i) = a_0 i^r + a_1 i^{r-1} + a_2 i^{r-2} \ldots + a_{r-1} i + a_r,$$

where $r \in \mathbf{N}$, $r, a_0, a_1, \ldots, a_{r-1}, a_r$ do not depend on i, and $a_0 \neq 0$. Obviously, we should have $r > 0$, $a_0 > 0$, and all coefficients $a_0, a_1, \ldots, a_{r-1}, a_r$ must be rational numbers. The function $2f(i) - f(i+1)$ is also a polynomial, namely

$$2f(i) - f(i+1) = a_0 i^r + b_1 i^{r-1} + b_2 i^{r-2} \ldots + b_{r-1} i + b_r$$

with the same a_0 and new coefficients $b_1, b_2, \ldots, b_{r-1}, b_r$ that are again rational numbers. Clearly, these new coefficients can be effectively found (assuming, of

course, that the degree r and the coefficients $a_0, a_1, \ldots, a_{r-1}, a_r$ are explicitly given or can be effectively found). Therefore, given any non-negative integer n_0, one can effectively find a positive integer i_0 satisfying Condition 1. This allows us to check algorithmically whether all elements of the set $[n_0 \,..\, \infty)$ are f–representable (the result can be obviously generalized to computable functions f such that $2f(i) - f(i+1)$ effectively diverges to $+\infty$ together with i, i.e. such that there is a computable function transforming any non-negative integer n_0 into some positive integer i_0 satisfying Condition 1).

Example 9 (a particular instance of Example 8). Let $f(i) = i^2$ for any i in $\mathbf{N}_+$. Then

$$2f(i) - f(i+1) = i^2 - 2i - 1 = i(i-2) - 1,$$

and therefore $2f(i) - f(i+1) \geq 129$ for any i in $[13..\infty)$. Since $129 + f(13) = 298$, the f–representability of all elements of $[129..\infty)$ is equivalent to the f–representability of the elements of $[129 \,..\, 298)$. The f–representability of the mentioned finitely many integers can be shown by computing the corresponding values of $f^\dagger$ (using Theorem 1) and showing that they are all positive, i.e. by a certain continuation of the computations that produced the table from Fig. 1. We have done this by computer, but we do not present the corresponding continuation of the table here. We preferred to present a table of f–representations of the numbers from 129 to 297 (cf. Fig. 2), since its correctness allows an easier manual verification (the table itself is produced by computer on the basis of Theorem 3; the representations are written without the curly brackets for the sake of saving space).

Remark. Some of the considered numbers have shorter f–representations than the ones given in the table. For instance, the number 131 has also the f–representation $\{1, 3, 11\}$. Note also that one could (especially at manual verification) use the remark after the proof of Theorem 4 and somewhat reduce the count of the numbers to be checked. In the concrete situation ($f(i) = i^2$, $n_0 = 129$) we have $S_i = [129 + i^2 \,..\, 2i^2)$ for any positive integer i. We see that $S_i = \emptyset$ for $i \leq 11$, $S_{12} = [273 \,..\, 288)$, and S_i consists of numbers not less than 298 for $i \geq 13$. Hence it would be enough to check the numbers belonging to $[129 \,..\, 298) \setminus S_{12}$, i.e. one could skip the check of 15 numbers.

Example 10 (several other particular instances of Example 8). Fig. 3 contains a summary of results of applying Theorem 4 to concrete polynomials f for obtaining results of the form "All elements of $[n_0 \,..\, \infty)$ are f–representable". In any of these results the number n_0 is the least possible for the polynomial in question and has been found by means of an iterative process starting with $n_0 = 0$ as an initial value. The iteration step and the termination of the process can be described as follows. We find a positive integer i_0 satisfying Condition 1 for the current n_0 and then we consecutively check for f-representability the numbers in $[n_0 \,..\, n_0 + f(i_0))$. If all of them turn out to be f–representable, then the process terminates with the current n_0 as its result. Otherwise, if m is the least number from $[n_0 \,..\, n_0 + f(i_0))$ that is not f–representable, then we take the number $m+1$ as a next value of n_0. Note that at the moment of the termination of the process all integers in the set $[0 \,..\, n_0 + f(i_0))$ turn out to have been already checked, hence the method can be obviously refined to compute also the total count of all positive integers that are not f–representable

n	f–representation of n	n	f–representation of n	n	f–representation of n	n	f–representation of n
129	4,7,8	172	1,4,5,7,9	215	3,6,7,11	258	5,8,13
130	7,9	173	4,6,11	216	4,6,8,10	259	5,7,8,11
131	3,4,5,9	174	5,7,10	217	6,9,10	260	8,14
132	1,3,4,5,9	175	3,6,7,9	218	7,13	261	6,15
133	4,6,9	176	1,3,6,7,9	219	5,7,8,9	262	4,5,10,11
134	3,5,10	177	4,5,6,10	220	3,4,5,7,11	263	5,6,9,11
135	3,4,5,6,7	178	3,13	221	10,11	264	3,5,7,9,10
136	6,10	179	3,7,11	222	4,6,7,11	265	11,12
137	4,11	180	6,12	223	2,5,7,8,9	266	8,9,11
138	5,7,8	181	9,10	224	4,8,12	267	4,7,9,11
139	3,7,9	182	5,6,11	225	15	268	3,5,7,8,11
140	2,6,10	183	3,5,7,10	226	4,5,8,11	269	10,13
141	4,5,10	184	2,6,12	227	5,9,11	270	7,10,11
142	5,6,9	185	8,11	228	3,5,7,8,9	271	4,5,7,9,10
143	2,3,7,9	186	4,7,11	229	6,7,12	272	4,16
144	12	187	2,3,5,7,10	230	7,9,10	273	4,7,8,12
145	8,9	188	1,2,3,5,7,10	231	5,6,7,11	274	7,15
146	5,11	189	5,8,10	232	6,14	275	5,9,13
147	3,5,7,8	190	4,5,7,10	233	8,13	276	5,7,9,11
148	2,12	191	5,6,7,9	234	7,8,11	277	9,14
149	7,10	192	1,5,6,7,9	235	4,5,7,8,9	278	3,10,13
150	3,4,5,10	193	7,12	236	3,5,9,11	279	5,6,7,13
151	3,5,6,9	194	7,8,9	237	4,10,11	280	6,10,12
152	4,6,10	195	5,7,11	238	6,9,11	281	6,8,9,10
153	3,12	196	14	239	3,7,9,10	282	7,8,13
154	4,5,7,8	197	4,9,10	240	3,5,6,7,11	283	3,7,15
155	5,7,9	198	4,5,6,11	241	4,15	284	3,5,9,13
156	2,4,6,10	199	3,4,5,7,10	242	5,6,9,10	285	8,10,11
157	6,11	200	6,8,10	243	5,7,13	286	6,9,13
158	4,5,6,9	201	4,8,11	244	10,12	287	6,7,9,11
159	2,5,7,9	202	9,11	245	8,9,10	288	3,5,6,7,13
160	4,12	203	3,7,8,9	246	5,10,11	289	17
161	5,6,10	204	3,5,7,11	247	4,5,6,7,11	290	11,13
162	4,5,11	205	6,13	248	4,6,14	291	5,8,9,11
163	3,4,5,7,8	206	6,7,11	249	6,7,8,10	292	6,16
164	8,10	207	4,5,6,7,9	250	9,13	293	7,10,12
165	4,7,10	208	8,12	251	7,9,11	294	7,8,9,10
166	6,7,9	209	4,7,12	252	3,5,7,13	295	5,7,10,11
167	3,4,5,6,9	210	5,8,11	253	3,10,12	296	10,14
168	2,8,10	211	4,5,7,11	254	6,7,13	297	4,6,8,9,10
169	13	212	4,14	255	5,7,9,10		
170	7,11	213	7,8,10	256	16		
171	4,5,7,9	214	3,6,13	257	7,8,12		

Fig. 2. Some f–representations of the numbers from 129 to 297 for $f(i) = i^2$

$f(i)$	n_0	i_0	$n_0 + f(i_0)$
$i(i+1)/2$	34	9	79
i^2+1	52	9	134
$(i+1)^2-1$	157	13	352
$i(i+1)(i+2)/6$	559	16	1375
i^3	12759	25	28384

Fig. 3. Several other instances of application of Theorem 4

(we established in this way the existence of exactly 2788 positive integers that are not f–representable in the case of $f(i) = i^3$). It is easy to design the process so that the output includes also the complete list of the non-representable positive integers.

Acknowledgments. Thanks are due to Professor S. Dodunekov and to Professor T. Tonkov for their valuable help in the search of sources discussing representation of natural numbers as sums of distinct squares.

REFERENCES

1. Halter-Koch, P. Darstellung natürlicher Zahlen als Summe von Quadraten. — Acta Arithmetica, **52**, 1982, 11–20.
2. Kassner, M. Darstellungen mit Nebenbedingungen durch quadratische Formen. — J. Reine Angew. Math., **331**, 1982, 151–161.
3. Lemoine, E. Questions 1793 et 1794. — Nouv. Ann. Math., 3, **17**, 1898, 195–196.
4. Pall, G. On sums of squares. — Amer. Math. Monthly, **40**, 1933, 10–18.
5. Ripert, M. L. Solution. — Nouv. Ann. Math., 3, **19**, 1900, 335–336.

Received on 04.07.1996

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

CONSTRUCTING MINIMAL PAIRS OF DEGREES*

IVAN N. SOSKOV

We prove that there exist sets of natural numbers A and B such that A and B form a minimal pair with respect to Turing reducibility, enumeration reducibility, hyperarithmetical reducibility and hyperenumeration reducibility. Relativized versions of this result are presented as well.

Keywords: Degrees, reducibilities, minimal pairs, forcing, enumerations.
1991/95 Mathematics Subject Classification: 03D30.

1. INTRODUCTION

In the present paper we consider four kinds of reducibilities among sets of natural numbers: Turing reducibility ($\leq_T$), enumeration reducibility ($\leq_e$), hyperarithmetical reducibility ($\leq_h$) and hyperenumeration reducibility ($\leq_{he}$). The first three of those reducibilities are well-known. The hyperenumeration reducibility has been introduced by Sanchis in [5] and further studied in [6]. It is a kind of positive reducibility which relates to hyperarithmetical reducibility, as enumeration reducibility relates to Turing reducibility.

Let $\sigma \in \{T, e, h, he\}$. By $\mathbf{0}_\sigma$ we shall denote the class

$$\{A \mid A \subseteq \mathbb{N} \ \& \ A \leq_\sigma \emptyset\}.$$

So, $\mathbf{0}_T$ consists of all recursive sets, $\mathbf{0}_e$ — of all recursively enumerable sets, $\mathbf{0}_h$ is equal to the class of all hyperarithmetical sets, and $\mathbf{0}_{he}$ consists of all Π^1_1 sets.

* Lecture presented at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.
This work was partially supported by the Ministry of Education, Science and Technologies, Contract I 412/95.

Two sets A and B are a minimal pair with respect to the σ-reducibility if for all sets X of natural numbers $X \leq_\sigma A \ \& \ X \leq_\sigma B \Rightarrow X \in \mathbf{0}_\sigma$.

It follows from the results of McEvoy and Cooper [3] that there exist sets of natural numbers A and B such that the pair (A, B) is minimal with respect to Turing reducibility and in the same time with respect to enumeration reducibility. Up to our knowledge, minimal pairs for the higher order reducibilities $\leq_h$ and $\leq_{he}$ are not well studied and an analogue of the result of McEvoy and Cooper is not known.

The aim of the present paper is to present a uniform construction of minimal pairs. In this way we shall obtain two sets A and B such that the pair (A, B) is minimal with respect to each of the reducibilities $\leq_T, \leq_e, \leq_h$ and $\leq_{he}$. Namely, we are going to prove the following theorem:

1.1. Theorem. *For every $A \subseteq \mathbb{N}$, such that $(\mathbb{N} \setminus A) \leq_e A$, there exists a $B \subseteq \mathbb{N}$ which is not Π^1_1 and such that if $\sigma \in \{T, e, h, he\}$, $X \leq_\sigma A$ and $X \leq_\sigma B$, then $X \in \mathbf{0}_\sigma$.*

In particular, if we pick up a sufficiently complex set A, i.e. if A is not Π^1_1, then we can find a set B such that for every $\sigma \in \{T, e, h, he\}$ the σ-degrees determined by the sets A and B form a minimal pair.

The proof of the theorem is based on a forcing technique introduced in [8] and used there for the purposes of the abstract recursion theory.

The paper is organized as follows. In Section 2 we summarize the basic definitions and results used in the sequel. In Section 3 we describe our forcing construction. The last Section 4 contains the proof of the theorem and some generalizations.

2. PRELIMINARIES

Throughout the paper we shall assume fixed a standard Gödel enumeration $W_0, \ldots, W_a, \ldots$ of the recursively enumerable sets. We shall assume also that an effective coding of the finite sets of natural numbers is given. By D_v we shall denote the finite set having code v.

By capital letters A, B, X etc. we shall denote sets of natural numbers.

We shall use the following definition of enumeration reducibility given in [4].

2.1. Definition. Let A and B be sets of natural numbers. Then A is *enumeration reducible* to B ($A \leq_e B$) if for some $a \in \mathbb{N}$ and for all $x \in \mathbb{N}$

$$x \in A \iff \exists v(\langle v, x \rangle \in W_a \ \& \ D_v \subseteq B).$$

Turing reducibility can be described in terms of enumeration reducibility. Given a set A, denote by A^+ the set $A \oplus (\mathbb{N} \setminus A)$. Then we have

$$A \leq_T B \iff A^+ \leq_e B^+.$$

Here $\oplus$ is the usual join operation. So,

$$x \in A \oplus B \iff \exists n((x = 2n \ \& \ n \in A) \vee (x = 2n + 1 \ \& \ n \in B)).$$

The notion of hyperenumeration reducibility is introduced in [5]. Let f, g denote arbitrary total functions in $\mathbb{N}$. By $\bar{f}(n)$ we shall denote (the code of) the sequence $\langle f(0), \ldots, f(n-1) \rangle$.

2.2. Definition. Given sets A and B of natural numbers, say that A is *hyperenumeration reducible* to B ($A \leq_{he} B$) if for some $a \in \mathbb{N}$ and for all $x \in \mathbb{N}$

$$x \in A \iff \forall f \exists n \exists v(\langle v, x, \bar{f}(n)\rangle \in W_a \ \&\ D_v \subseteq B).$$

From the definition it follows immediately that A is Π^1_1 in B iff $A \leq_{he} B^+$ and hence we can express hyperarithmetical reducibility in terms of hyperenumeration reducibility:

$$A \leq_h B \iff A^+ \leq_{he} B^+.$$

A set A of natural numbers is called *total* if $(\mathbb{N} \setminus A) \leq_e A$ or, equivalently, if $A^+ \leq_e A$. The following obvious lemma shows that if two total sets form a minimal pair with respect to enumeration reducibility and hyperenumeration reducibility, then they form a minimal pair with respect to Turing reducibility and with respect to hyperarithmetical reducibility.

2.3. Lemma. *Let A and B be total sets of natural numbers. Then:*
(i) $\forall X(X \leq_e A \ \&\ X \leq_e B \Rightarrow X \in \mathbf{0}_e) \Rightarrow \forall X(X \leq_T A \ \&\ X \leq_T B \Rightarrow X \in \mathbf{0}_T)$;
(ii) $\forall X(X \leq_{he} A \ \&\ X \leq_{he} B \Rightarrow X \in \mathbf{0}_{he}) \Rightarrow \forall X(X \leq_h A \ \&\ X \leq_h B \Rightarrow X \in \mathbf{0}_h)$.

We shall identify the partial predicates on $\mathbb{N}$ with the partial functions, taking values in $\{0, 1\}$, assuming that 0 stands for true and 1 for false.

By $\mathfrak{A}_\Sigma$ we shall denote the structure $(\mathbb{N}; G, \Sigma)$, where G is a total binary predicate which is equal to the graph of the successor function, in other words,

$$G(x, y) \simeq \begin{cases} 0, & \text{if } y = x + 1, \\ 1 & \text{otherwise,} \end{cases}$$

and Σ is a unary partial predicate on the natural numbers.

Enumeration of $\mathfrak{A}_\Sigma$ is a total surjective mapping f of $\mathbb{N}$ onto $\mathbb{N}$. Clearly, every enumeration determines a unique structure $\mathfrak{B}_f = (\mathbb{N}; G^{\mathfrak{B}_f}, \Sigma^{\mathfrak{B}_f})$, where for all x, y

$$G^{\mathfrak{B}_f}(x, y) \simeq G(f(x), f(y)) \quad \text{and} \quad \Sigma^{\mathfrak{B}_f}(x) \simeq \Sigma(f(x)).$$

Given an enumeration f of $\mathfrak{A}_\Sigma$, denote by $D(\mathfrak{B}_f)$ the set of all Gödel numbers of the elements of the diagram of $\mathfrak{B}_f$. In other words,

$$D(\mathfrak{B}_f) = \{\langle 1, n, m, \varepsilon\rangle \mid G^{\mathfrak{B}_f}(n, m) \simeq \varepsilon\} \cup \{\langle 2, n, \varepsilon\rangle \mid \Sigma^{\mathfrak{B}_f}(n) \simeq \varepsilon\}.$$

Notice that if the predicate Σ is total, then $D(\mathfrak{B}_f)$ is a total set.

The main property of the structure $\mathfrak{A}_\Sigma$ is that it is relatively stable. This means that for every enumeration f of $\mathfrak{A}_\Sigma$ the function f is partial recursive relatively $D(\mathfrak{B}_f)$, i.e. $graph(f) \leq_e D(\mathfrak{B}_f)$.

2.4. Proposition. *Let f be an enumeration of $\mathfrak{A}_\Sigma$. Then f is partial recursive in $D(\mathfrak{B}_f)$.*

Proof. Let us fix a natural number 0_f such that $f(0_f) = 0$. First we are going to show that

$$f(n) = 0 \iff \exists y \left(G^{\mathfrak{B}_f}(0_f, y) \ \&\ G^{\mathfrak{B}_f}(n, y)\right).$$

Indeed, suppose that $f(n) = 0$. Take an y such that $f(y) = 1$. Then we have $G(f(0_f), f(y))$ and $G(f(n), f(y))$, and hence $G^{\mathfrak{B}_f}(0_f, y)$ and $G^{\mathfrak{B}_f}(n, y)$. Now

suppose that for some y, $G^{\mathfrak{B}_f}(0_f, y)$ and $G^{\mathfrak{B}_f}(n, y)$. Then $f(y) = 1$ and since $G(f(n), 1)$, we get that $f(n) = 0$.

In the same way one can show for $k > 0$ that

$$f(n) = k \iff \exists x_1 \ldots x_{k-1} \left(G^{\mathfrak{B}_f}(0_f, x_1) \,\&\, \ldots \,\&\, G^{\mathfrak{B}_f}(x_{k-2}, x_{k-1}) \,\&\, G^{\mathfrak{B}_f}(x_{k-1}, n)\right).$$

So the graph of f is enumeration reducible to $D(\mathfrak{B}_f)$ and hence f is partial recursive in $D(\mathfrak{B}_f)$. ∎

2.5. Corollary. *For every enumeration f of $\mathfrak{A}_\Sigma$, $\Sigma \leq_e D(\mathfrak{B}_f)$.*

2.6. Definition. Let $A \subseteq \mathbb{N}$, $\sigma \in \{T, e, h, he\}$ and f be an enumeration of $\mathfrak{A}_\Sigma$. Then A is *σ-admissible* in f if $f^{-1}(A) \leq_\sigma D(\mathfrak{B}_f)$.

Now we are ready to describe the plan of the proof of Theorem 1.1. Let Σ be a total recursive predicate, for example let $\Sigma = \lambda x.0$.

Given a total set A, denote by Q_σ, $\sigma \in \{e, he\}$, the class of all sets which are σ-reducible to A. In what follows we shall show that there exists an enumeration f of $\mathfrak{A}_\Sigma$ having the following properties:

(1) f and hence $D(\mathfrak{B}_f)$ is not Π^1_1;

(2) If $\sigma \in \{e, he\}$, $X \in Q_\sigma$ and X is σ-admissible in f, then $X \in \mathbf{0}_\sigma$.

Denote the set $D(\mathfrak{B}_f)$ by B. Now suppose that $\sigma \in \{e, he\}$ and $X \leq_\sigma A$ and $X \leq_\sigma B$. Using the stability of $\mathfrak{A}_\Sigma$, we obtain from here that X is σ-admissible in f and hence, by (2), $X \in \mathbf{0}_\sigma$.

From here by Lemma 2.3 we obtain for all $\sigma \in \{T, e, h, he\}$

$$X \leq_\sigma A \,\&\, X \leq_\sigma B \Rightarrow X \in \mathbf{0}_\sigma.$$

In the same way, using appropriate definitions of the predicate Σ, we shall obtain also relativized versions of the theorem.

3. GENERIC ENUMERATIONS

Every finite mapping of $\mathbb{N}$ into $\mathbb{N}$ is called *finite part*. By Δ we shall denote the set of all finite parts. Elements of Δ will be denoted by lowercase Greek letters $\delta, \tau, \rho, \ldots$. We shall use "$\subseteq$" to denote the usual inclusion relation on partial functions. Clearly, "$\subseteq$" induces a partial ordering on Δ.

3.1. Definition. Let $E \subseteq \Delta$ and f be an enumeration of $\mathfrak{A}_\Sigma$. Then:

(1) E is *dense* if for every $\delta \in \Delta$ there exists a $\tau \in E$ such that $\delta \subseteq \tau$;

(2) E is *dense in the enumeration f* if for every finite part $\delta \subseteq f$ there exists a $\tau \in E$ such that $\delta \subseteq \tau$;

(3) f *meets* E if there exists a finite part $\delta \in E$ such that $\delta \subseteq f$.

Notice that a dense set E is automatically dense in every enumeration of $\mathfrak{A}_\Sigma$. Let $\mathcal{F}$ be a countable family of subsets of Δ.

3.2. Definition. An enumeration f is *$\mathcal{F}$-generic* if

$$(\forall E \in \mathcal{F})(E \text{ is dense in } f \Rightarrow f \text{ meets } E).$$

Let $D(\Sigma) = \{\langle n, \varepsilon\rangle \mid \Sigma(n) \simeq \varepsilon\}$. Let $\sigma \in \{e, he\}$. Given a set A, say that $A \leq_\sigma \Sigma$ if $A \leq_\sigma D(\Sigma)$. For a function f let $f \leq_\sigma \Sigma$ if $graph(f) \leq_\sigma D(\Sigma)$.

3.3. Proposition. *Let $\delta \in \Delta$. There exists an $\mathcal{F}$-generic enumeration f of $\mathfrak{A}_\Sigma$ which extends δ and such that $f \not\leq_{he} \Sigma$.*

Proof. A usual finite end-extension construction of the mapping f. Start with $\delta_0 = \delta$. Consider three kinds of steps. On steps $q = 3r$ ensure that f is total and surjective. On steps $q = 3r + 1$ ensure the genericity. Finally, on steps $q = 3r + 2$ consider the r-th he-reducible to Σ partial function ψ_r and ensure that $f \neq \psi_r$. ■

Denote by $\mathcal{E}$ the class of all enumerations of $\mathfrak{A}_\Sigma$.

3.4. Definition. Let $S \subseteq \mathbb{N} \times \mathcal{E}$. The set S is called *complete* relative to $\mathcal{F}$ if for every $n \in \mathbb{N}$, $\delta \in \Delta$ there exists a $\tau \supseteq \delta$ such that if f is $\mathcal{F}$-generic and $\tau \subseteq f$, then the pair (n, f) belongs to S.

The next proposition is a generalized version of Proposition 3.7 [8]. The simple proof presented here is based on a suggestion of Vl. Soskov.

3.5. Proposition. *Let $S \subseteq \mathbb{N} \times \mathcal{E}$ be complete relative to $\mathcal{F}$. Then there exists a countable family $\mathcal{F}_S$ of subsets of Δ such that if f is $\mathcal{F}_S$-generic, then $\forall n((n, f) \in S)$.*

Proof. Given a natural number n, let

$$E_n = \{\tau \mid \forall f(f \text{ is } \mathcal{F}\text{-generic } \& \ \tau \subseteq f \Rightarrow (n, f) \in S)\}.$$

It follows from the completeness of S that the set E_n is dense.

Denote by $\mathcal{F}_S$ the family $\{E_n \mid n \in \mathbb{N}\} \cup \mathcal{F}$. Suppose that f is $\mathcal{F}_S$-generic. Fix an $n \in \mathbb{N}$. Since E_n is dense, f meets it. Let $\tau \in E_n$ be such that $\tau \subseteq f$. Clearly, f is $\mathcal{F}$-generic. Hence, by the definition of E_n, $(n, f) \in S$. ■

Let $\sigma \in \{e, he\}$ and let $P_0^\sigma, \ldots, P_a^\sigma, \ldots$ be a sequence of unary predicate letters. Assume that a satisfaction relation "$f \vDash_\sigma P_a^\sigma(x)$" is defined, so that for every enumeration f of $\mathfrak{A}_\Sigma$

$$A \leq_\sigma D(\mathfrak{B}_f) \iff \exists a(A = \{x \mid f \vDash_\sigma P_a^\sigma(x)\}).$$

Suppose also that "$\delta \Vdash_\sigma P_a^\sigma(x)$" is a forcing relation satisfying the following *forcing conditions*:

(F1) $\delta \subseteq \tau \ \& \ \delta \Vdash_\sigma P_a^\sigma(x) \Rightarrow \tau \Vdash_\sigma P_a^\sigma(x)$;

(F2) There exists a countable family $\mathcal{F}_\sigma$ of subsets of Δ such that for every $\mathcal{F}_\sigma$-generic enumeration f, $f \vDash_\sigma P_a^\sigma(x) \iff (\exists \delta \subseteq f)(\delta \Vdash_\sigma P_a^\sigma(x))$..

3.6. Definition. Let $A \subseteq \mathbb{N}$. The set A has a *σ-normal form* if for some $a \in \mathbb{N}$, $\delta \in \Delta$ and for all $n \notin dom(\delta)$, $x \in \mathbb{N}$,

$$x \in A \iff \exists \tau(\delta \subseteq \tau)(\tau(n) \simeq x \ \& \ \tau \Vdash_\sigma P_a^\sigma(n)). \tag{3.1}$$

Given a set A, call P_a^σ an *f-associate* of A if for all $n \in \mathbb{N}$

$$f(n) \in A \iff f \vDash_\sigma P_a^\sigma(n).$$

Assume that the recursive pairing function $\langle \cdot, \cdot \rangle$ is chosen, so that every natural number is a code of an ordered pair.

3.7. Proposition. *Let $Q = \{A_0, A_1, \ldots, A_r, \ldots\}$ be a countable family of subsets of $\mathbb{N}$. Let the subset S of $\mathbb{N} \times \mathcal{E}$ be defined by*

$$(\langle a, r\rangle, f) \in S \iff A_r \text{ has a } \sigma\text{-normal form or } P_a^\sigma \text{ is not an } f\text{-associate of } A_r.$$

Then S is complete relative to $\mathcal{F}_\sigma$.

Proof. Let us fix a natural number $m = \langle a, r\rangle$ and a finite part δ. Assume that A_r has a σ-normal form. Clearly, for every enumeration f the pair (m, f) belongs to S.

Now suppose that A_r does not have a σ-normal form. Then there exist natural numbers x and $n \notin dom(\delta)$ for which the equivalence (3.1) fails. We have two possibilities. First suppose that

$$x \in A \;\&\; \forall \tau(\delta \subseteq \tau)(\tau(n) \simeq x \Rightarrow \tau \nVdash_\sigma P_a^\sigma(n)).$$

Take a τ such that $\delta \subseteq \tau \;\&\; \tau(n) \simeq x$. Let f be an $\mathcal{F}_\sigma$-generic enumeration which extends τ. Clearly, $f(n) = x \in A_r$. Assume that $f \vDash_\sigma P_a^\sigma(n)$. Then, by (F2), there exists a $\rho \subseteq f$ such that $\rho \Vdash_\sigma P_a^\sigma(n)$. By (F1) we may assume that $\tau \subseteq \rho$. A contradiction. So, P_a^σ is not an f-associate of A_r and hence $(m, f) \in S$.

Now suppose that

$$x \notin A_r \;\&\; \exists \tau(\delta \subseteq \tau)(\tau(n) \simeq x \;\&\; \tau \Vdash_\sigma P_a^\sigma(n)).$$

Let f be $\mathcal{F}_\sigma$-generic and $\tau \subseteq f$. Then, by (F2), $f \vDash_\sigma P_a^\sigma(n)$ but $f(n) = x \notin A_r$. Hence $(m, f) \in S$. ∎

Combining the last proposition and Proposition 3.5, we get the following

3.8. Corollary. *Let Q be a countable family of sets of natural numbers. There exists a countable family $\mathcal{F}$ of subsets of Δ such that if f is $\mathcal{F}$-generic, $A \in Q$ and A is σ-admissible in f, then A has a σ-normal form.*

4. PROOF OF THE THEOREM

We start by defining appropriate $\vDash_\sigma$ and $\Vdash_\sigma$ relations for $\sigma \in \{e, he\}$. Consider first $\sigma = e$.

4.1. Definition. Given natural number $a \in N$ and enumeration f of $\mathfrak{A}_\Sigma$, let

$$f \vDash_e P_a^e(n) \iff \exists v(\langle v, n\rangle \in W_a \;\&\; D_v \subseteq D(\mathfrak{B}_f)).$$

From the definition above it follows immediately that for every enumeration f and $A \subseteq \mathbb{N}$

$$A \leq_e D(\mathfrak{B}_f) \iff \exists a(A = \{n \mid f \vDash P_a^e(n)\}). \tag{4.1}$$

The definition of the forcing relation $\Vdash_e$ is a little bit more complicated. Let δ be finite part. Given a natural number u, let $\delta \Vdash_e u$ if $u = \langle 1, n, m, \varepsilon\rangle$ for some n, m in $dom(\delta)$ and $G(\delta(n), \delta(m)) \simeq \varepsilon$ or $u = \langle 2, n, \varepsilon\rangle$ for some $n \in dom(\delta)$ and $\Sigma(\delta(n)) \simeq \varepsilon$.

For a finite set D let $\delta \Vdash_e D \iff (\forall u \in D)(\delta \Vdash_e u)$.

Finally, given $a \in \mathbb{N}$, let

$$\delta \Vdash_e P_a^e(n) \iff \exists v(\langle v, n\rangle \in W_a \;\&\; \delta \Vdash_e D_v).$$

It is obvious that the forcing conditions (F1) and (F2) hold for $\vDash_e$ and $\Vdash_e$, where the family $\mathcal{F}_e$ is empty.

4.2. Proposition. *Let $A \subseteq \mathbb{N}$ have an e-normal form. Then $A \leq_e \Sigma$.*

Proof. Let δ and a be such that (3.1) holds for all $n \notin dom(\delta)$ and $x \in \mathbb{N}$. Fix an $n_0 \notin dom(\delta)$. Then

$$x \in A \iff \exists \tau(\delta \subseteq \tau)(\tau(n_0) \simeq x \ \& \ \tau \Vdash_e P_a^e(n_0)).$$

Assume that an effective coding of the finite parts is fixed. From the definition of $\Vdash_e$, using the recursiveness of G, we obtain that the set $\{\tau \mid \tau \Vdash_e P_a^e(n_0)\}$ is e-reducible to Σ. Therefore $A \leq_e \Sigma$. ∎

Now let us turn to the hyperenumeration case. Consider two sequences

$$R_0, \ldots, R_a, \ldots; \quad F_0, \ldots, F_a, \ldots$$

of new binary predicate letters. Given an enumeration f, let

$$f \vDash_{he} R_a(x, s) \iff \exists v(\langle v, x, s\rangle \in W_a \ \& \ D_v \subseteq D(\mathfrak{B}_f)).$$

Let s denote (codes of) arbitrary finite strings of natural numbers. If $s = \langle z_1, \ldots, z_n\rangle$, then by $s * z$ we shall denote the string $\langle z_1, \ldots, z_n, z\rangle$. By $\langle\rangle$ we shall denote the empty string.

Given a finite string s and a natural number x, define $f \vDash_{he} F_a(x, s)$ by means of the following inductive

4.3. Definition.

If $f \vDash_{he} R_a(x, s)$, then $f \vDash_{he} F_a(x, s)$;

If $\forall z(f \vDash_{he} F_a(x, s * z))$, then $f \vDash_{he} F_a(x, s)$.

Suppose that $f \vDash F_a(x, s)$. By $|x, s|$ we shall denote the first ordinal at which the pair (x, s) appears in the inductive definition. In other words,

$$|x, s| = \begin{cases} 0, & \text{if } f \vDash_{he} R_a(x, s), \\ \sup(|x, s * z| + 1 : z \in \mathbb{N}) & \text{otherwise.} \end{cases}$$

4.4. Lemma. *Let $A \subseteq \mathbb{N}$ and f be an enumeration of $\mathfrak{A}_\Sigma$. Then*

$$A \leq_{he} D(\mathfrak{B}_f) \iff \exists a(A = \{x \mid f \vDash_{he} F_a(x, \langle\rangle)\}).$$

Proof. By definition $A \leq_{he} D(\mathfrak{B}_f)$ if, and only if, for some $a \in \mathbb{N}$

$$x \in A \iff \forall g \exists n \exists v(\langle v, x, \bar{g}(n)\rangle \in W_a \ \& \ D_v \subseteq D(\mathfrak{B}_f)).$$

Hence $A \leq_{he} D(\mathfrak{B}_f)$ iff there exists $a \in \mathbb{N}$ such that

$$x \in A \iff \forall g \exists n(f \vDash_{he} R_a(x, \bar{g}(n))).$$

We shall show that

$$\forall g \exists n(f \vDash_{he} R_a(x, \bar{g}(n))) \iff f \vDash_{he} F_a(x, \langle\rangle). \tag{4.2}$$

Suppose that the left hand side of (4.2) holds. Towards a contradiction assume that $f \nvDash_{he} F_a(x, \langle\rangle)$. Then there exists a sequence $z_0, z_1, \ldots, z_n, \ldots$ of natural numbers such that if $s_n = \langle z_0, \ldots, z_{n-1}\rangle$, then

$$f \nvDash_{he} R_a(x, s_n) \ \& \ f \nvDash_{he} F_a(s_n * z_n, x). \tag{4.3}$$

The construction of $z_0, z_1, \ldots, z_n, \ldots$ is by induction on n. Since $f \not\models_{he} F_a(x, \langle\rangle)$, $f \not\models_{he} R_a(x, \langle\rangle)$ and for some z, $f \not\models_{he} F_e(x, \langle z\rangle)$. Set $z_0 = z$.

Suppose that $z_0, \ldots, z_n$ are chosen, so that (4.3) holds. Let $s_{n+1} = \langle z_0, \ldots, z_n\rangle$. By (4.3) $f \not\models_{he} R_a(x, s_{n+1})$ and for some z, $f \not\models_{he} F_a(x, s_{n+1} * z)$. Take $z_{n+1} = z$.

Now let $g(n) = z_n$. Clearly, $\forall n(f \not\models_{he} R_a(x, \bar{g}(n)))$.

Given a finite string $s = \langle z_0, \ldots, z_{n-1}\rangle$ and a function g, let

$$s \subseteq g \iff (\forall k < n)(g(k) = z_k).$$

To prove (4.2) in the right to left direction, we shall show by means of transfinite induction on $|x, s|$ that

$$f \models_{he} F_a(x, s) \Rightarrow \forall g \supseteq s \exists n(f \models R_a(x, \bar{g}(n))) \tag{4.4}$$

and use that every function extends the empty string $\langle\rangle$.

Indeed, if $f \models_{he} R_a(x, s)$, then (4.4) is obvious. Suppose that $f \not\models_{he} R_a(x, s)$. By induction $(\forall z)(\forall g \supseteq s * z)\exists n(f \models_{he} R_a(x, \bar{g}(n)))$. Suppose that $g \supseteq s$. Then for some z, $g \supseteq s * z$ and hence $\exists n(f \models_{he} R_a(x, \bar{g}(n)))$. ■

Let $f \models_{he} P_a^{he}(x) \iff f \models_{he} F_a(x, \langle\rangle)$.

Our next task is to define an appropriate forcing relation $\delta \Vdash_{he} P_a^{he}(x)$. First let

$$\delta \Vdash_{he} R_a(x, s) \iff \exists v(\langle v, x, s\rangle \in W_a \ \&\ \delta \Vdash_e D_v).$$

Clearly, we have as for enumeration reducibility:

(R1) $\delta \Vdash_{he} R_a(x, s)$ & $\delta \subseteq \tau \Rightarrow \tau \Vdash_{he} R_a(x, s)$;

(R2) For every enumeration f, $f \models_{he} R_a(x, s) \iff \exists \delta \subseteq f(\delta \Vdash_{he} R_a(x, s))$.

Now we are ready to define $\delta \Vdash_{he} F_a(x, s)$ by means of the following inductive definition.

4.5. Definition.

If $\delta \Vdash_{he} R_a(x, s)$, then $\delta \Vdash_{he} F_a(x, s)$;

If $\forall z \in N \forall \tau \supseteq \delta \exists \rho \supseteq \tau(\rho \Vdash_{he} F_a(x, s * z))$, then $\delta \Vdash_{he} F_a(x, s)$.

We associate ordinals with the tuples (δ, x, s) such that $\delta \Vdash_{he} F_a(x, s)$ as usual:

$$|\delta, x, s| = \begin{cases} 0, & \text{if } \delta \Vdash_{he} R_a(x, s), \\ \sup(\min(|\rho, x, s * z| + 1 : \rho \supseteq \tau) : \tau \supseteq \delta, z \in \mathbb{N}) & \text{otherwise.} \end{cases}$$

The next lemma follows immediately from Definition 4.5.

4.6. Lemma. *Let δ, τ be finite parts, $\delta \subseteq \tau$ and $\delta \Vdash_{he} F_a(x, s)$, then $\tau \Vdash_{he} F_a(x, s)$.*

Let $\mathcal{F}_1$ be the family of all subsets

$$E_{\delta,x,s,z} = \{\rho \mid \rho \Vdash_{he} F_a(x, s * z) \ \&\ |\rho, x, s * z| < |\delta, x, s|\} \text{ of } \Delta.$$

4.7. Lemma. *Let f be an $\mathcal{F}_1$-generic enumeration, $\delta \subseteq f$ and $\delta \Vdash_{he} F_a(x, s)$. Then $f \models_{he} F_a(x, s)$.*

Proof. Transfinite induction on $|\delta, x, s|$. Skipping the obvious case $f \vDash_{he} R_a(x,s)$, assume $f \nvDash_{he} R_a(x,s)$. Fix a $z \in \mathbb{N}$ and consider the element

$$E = \{\rho \mid \rho \Vdash_{he} F_a(x, s*z) \ \&\ |\rho, x, s*z| < |\delta, x, s|\}$$

of $\mathcal{F}_1$. We shall show that E is dense in f. Let $\mu \subseteq f$. Take a $\tau \subseteq f$ such that $\mu \subseteq \tau$ and $\delta \subseteq \tau$. Since $f \nvDash_{he} R_a(x,s)$, by (R2), $\delta \nVdash_{he} R_a(x,s)$ and hence, by Definition 4.5, there exists a $\rho \supseteq \tau$ which belongs to E.

From here, by genericity, there exists a $\rho \subseteq f$ which belongs to E.

Now we have that $|\rho, x, s*z| < |\delta, x, s|$ and $\rho \Vdash_{he} F_a(x, s*z)$. Hence, by the induction hypothesis, $f \vDash F_a(x, s*z)$.

So we have proved that $\forall z(f \vDash F_a(x, s*z))$, and hence $f \vDash_{he} F_a(x,s)$. ∎

Denote by $\mathcal{F}_2$ the family containing all sets $\{\tau \mid \exists z \forall \rho \supseteq \tau(\rho \nVdash_{he} F_a(x, s*z))\}$.

4.8. Lemma. *Let f be $\mathcal{F}_2$-generic and $f \vDash_{he} F_a(x,s)$. Then there exists a $\delta \subseteq f$ such that $\delta \Vdash_{he} F_a(x,s)$.*

Proof. Transfinite induction on $|x,s|$. Assume that $\forall \delta \subseteq f(\delta \nVdash_{he} F_a(x,s))$. Then the set $E = \{\tau \mid \exists z \forall \rho \supseteq \tau(\rho \nVdash_{he} F_a(x, s*z))\}$ is dense in f. By genericity, there exist $\tau \subseteq f$ and $z \in \mathbb{N}$, such that $\forall \rho \supseteq \tau(\rho \nVdash_{he} F_a(x, s*z))$.

On the other hand, $f \vDash_{he} F_a(x,s)$ and $f \nvDash_{he} R_a(x,s)$. (Otherwise we could find a $\delta \subseteq f$ such that $\delta \Vdash_{he} R_a(x,s)$.) Therefore $f \vDash F_a(x, s*z)$, and hence, by induction, there exists a $\rho \subseteq f$ such that $\rho \Vdash F_a(x, s*z)$. By Lemma 4.6 we may assume that $\tau \subseteq \rho$. A contradiction. ∎

Define $\delta \Vdash_{he} P_a^{he}(x) \iff \delta \Vdash_{he} F_a(x, \langle\rangle)$.

Let $\mathcal{F}_{he} = \mathcal{F}_1 \cup \mathcal{F}_2$. Combining the last three lemmas we obtain that $\vDash_{he}$ and $\Vdash_{he}$ satisfy the forcing conditions (F1) and (F2).

4.9. Proposition. *Suppose that A has a he-normal form. Then $A \leq_{he} \Sigma$.*

Proof. Let δ and a be such that for all $n \notin dom(\delta)$ and x

$$x \in A \iff \exists \tau \supseteq \delta(\tau(n) \simeq x \ \&\ \tau \Vdash_{he} F_a(n, \langle\rangle)).$$

Consider the set $P = \{(\tau, n, s) \mid \tau \Vdash_{he} F_a(n,s)\}$. We are going to show that $P \leq_{he} \Sigma$. For this purpose we shall give a game characterization of the forcing "$\Vdash_{he}$". Our game starts over a triple (τ, n, s) and has two players — $(\forall)$ and $(\exists)$. If $\tau \Vdash_{he} R_a(n,s)$, then the game stops and $(\exists)$ wins. Otherwise the first player $(\forall)$ chooses a natural number z and a finite part $\mu \supseteq \tau$. Then the second player $(\exists)$ chooses a finite part $\nu \supseteq \mu$. The game continues over $(\nu, n, s*z)$. Now our claim is that $\tau \Vdash_{he} F_a(n,s)$ iff there exists a strategy for $(\exists)$ for winning every game over (τ, n, s). To formulate this claim precisely, we shall represent the possible moves of $(\forall)$ by two total functions g_1 and g_2, where $g_1(\tau, n, s)$ will choose the natural number z and $g_2(\tau, n, s)$ will give the finite part μ. We shall call the pair (g_1, g_2) correct if $\forall \tau \forall n \forall s(\tau \subseteq g_2(\tau, n, s))$.

4.10. Claim. *$\tau \Vdash_{he} F_a(n,s)$ iff for every correct pair (g_1, g_2) there exists a finite nonempty sequence $\langle \nu_0, \nu_1, \ldots, \nu_k \rangle$ of finite parts such that if*

$$z_1 = g_1(\nu_0, n, s), z_2 = g_1(\nu_1, n, s*z_1), \ldots, z_k = g_1(\nu_{k-1}, n, s*z_1*\ldots*z_{k-1}),$$

then:

a) $\tau = \nu_0$;

b) $(\forall i < k)(g_2(\nu_i, n, s * z_1 * \ldots * z_i) \subseteq \nu_{i+1})$;

c) $\nu_k \Vdash_{he} R_a(n, s * z_1 * \ldots * z_k)$.

Proof. The proof of the left to right direction is by induction on $|\tau, n, s|$. Suppose that $\tau \Vdash_{he} F_a(n, s)$. Let (g_1, g_2) be a correct pair of functions. If $\tau \Vdash_{he} R_a(n, s)$, then the sequence $\langle \tau \rangle$ satisfies the conditions a)–c). Suppose now that $\tau \nVdash_{he} R_a(n, s)$. Let $z_1 = g_1(\tau, n, s)$ and $\mu = g_2(\tau, n, s)$. By the correctness of (g_1, g_2), $\tau \subseteq \mu$. By the definition of $\Vdash_{he}$ there exists a $\nu_1 \supseteq \mu$ such that $\nu_1 \Vdash_{he} F_a(n, s * z_1)$ and $|\nu_1, n, s * z_1| < |\tau, n, s|$. By induction there exists a finite non-empty sequence $\langle \nu_1, \ldots, \nu_k \rangle$ of finite parts, satisfying the conditions a)–c) with respect to $(\nu_1, n, s * z_1)$. Now it is trivial to show that the sequence $\langle \tau, \nu_1, \ldots, \nu_k \rangle$ satisfies a)–c) with respect to (τ, n, s).

Suppose now that $\tau \nVdash_{he} F_a(n, s)$. We shall show that there exists a correct pair (g_1, g_2) of functions for which there is no finite sequence of finite parts satisfying a)–c). Given finite part δ and string t, check if there exist z and $\mu \supseteq \delta$ such that $(\forall \nu \supseteq \mu)(\nu \nVdash_{he} F_a(n, t * z))$. In case of a positive answer let $g_1(\delta, n, t)$ be one of those z and $g_2(\delta, n, t)$ be one of those μ. If the answer is negative, then let $g_1(\delta, n, t) = 0$ and $g_2(\delta, n, t) = \delta$. Clearly, the pair (g_1, g_2) is correct.

Now assume that $\langle \nu_0, \ldots, \nu_k \rangle$ is a sequence of finite parts satisfying the conditions a)–c). By a) we have $\nu_0 = \tau$. Since $\nu_0 \nVdash_{he} F_a(n, s)$, $\nu_0 \nVdash_{he} R_a(n, s)$, and

$$\exists z \exists \mu \supseteq \nu_0 \forall \nu \supseteq \mu (\nu \nVdash_{he} F_a(n, s * z)).$$

By the definition of g_1 and g_2 and b) we get $\nu_1 \nVdash_{he} F_a(n, s * z_1)$. So, proceeding as above, we have that

$$\nu_1 \nVdash_{he} R_a(n, s * z_1),\ \nu_2 \nVdash_{he} R_a(n, s * z_1 * z_2),\ \ldots,\ \nu_k \nVdash_{he} R_a(n, s * z_1 * \ldots * z_k).$$

The last contradicts c). ∎

Using the Claim and the fact that the set $\{(\tau, n, s) \mid \tau \Vdash_{he} R_a(n, s)\}$ is enumeration reducible to Σ, we obtain immediately that $P \leq_{he} \Sigma$ and hence that $A \leq_{he} \Sigma$. ∎

Now we are ready to prove the main results.

4.11. Theorem. *Let C and A be total sets. There exists a total set B such that $C \leq_T B$, $B \not\leq_{he} C$ and for all $\sigma \in \{T, e, h, he\}$ and all $X \subseteq \mathbb{N}$*

$$X \leq_\sigma A \ \&\ X \leq_\sigma B \Rightarrow X \leq_\sigma C.$$

Proof. Let

$$\Sigma(x) = \begin{cases} 0, & \text{if } x \in C, \\ 1 & \text{otherwise.} \end{cases}$$

Since C is total, we have for all $\sigma \in \{T, e, h, he\}$ that $C \leq_\sigma \Sigma$ and $\Sigma \leq_\sigma C$, i. e. $C \equiv_\sigma \Sigma$.

Let A be a total set. Denote by Q_σ, $\sigma \in \{e, he\}$, the family of all sets which are σ-reducible to A. By Corollary 3.8 there exist denumerable families $\mathcal{F}_{Q_\sigma}$ of subsets of Δ such that if f is $\mathcal{F}_{Q_\sigma}$-generic, $X \in Q_\sigma$ and X is σ-admissible in f, then X has a σ-normal form. Let f be an enumeration of $\mathfrak{A}_\Sigma$ which is not he-reducible to Σ

and generic with respect to $\mathcal{F}_{Q_e} \cup \mathcal{F}_{Q_{he}}$. Denote $D(\mathfrak{B}_f)$ by B. Since the predicate Σ is totally defined, the set B is total. By the stability of $\mathfrak{A}_\Sigma$, $f \leq_{he} B$ and hence $B \not\leq_{he} \Sigma$ and $\Sigma \leq_T B$.

By Lemma 2.3 it is sufficient to show for $\sigma \in \{e, he\}$

$$X \leq_\sigma A \ \& \ X \leq_\sigma B \Rightarrow X \leq_\sigma C.$$

Now suppose that $X \leq_\sigma A$ and $X \leq_\sigma B$. Since f is partial recursive in B, $f^{-1}(X) \leq_\sigma B$. So $X \in Q_\sigma$ and X is σ-admissible in f. From here it follows that X has a σ-normal form and hence by Proposition 4.2 and Proposition 4.9, respectively, $X \leq_\sigma \Sigma$. Therefore $X \leq_\sigma C$. ∎

Notice that since $\emptyset$ is total, Theorem 1.1 is a direct corollary of the above theorem.

If we start by an arbitrary, not necessarily total set C, then we can prove a similar result but only for the positive reducibilities $\leq_e$ and $\leq_{he}$.

4.12. Theorem. *Let C and A be subsets of $\mathbb{N}$. There exists a subset B of $\mathbb{N}$ such that $C \leq_e B$, $B \not\leq_{he} C$ and if $\sigma \in \{e, he\}$, then for all $X \subseteq \mathbb{N}$*

$$X \leq_\sigma A \ \& \ X \leq_\sigma B \Rightarrow X \leq_\sigma C.$$

Proof. Let us define the partial predicate Σ by

$$\Sigma(x) = \begin{cases} 0, & \text{if } x \in C, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Now we have for $\sigma \in \{e, he\}$ that $\Sigma \equiv_\sigma C$. From here the theorem follows by an almost literal repeating of the arguments used in the proof of the previous theorem. ∎

The method used in the proofs of the theorems above allows further generalizations and applications. We may add countably many satisfaction and forcing relations to the so far considered $\models_\sigma$ and $\Vdash_\sigma$, $\sigma \in \{e, he\}$, relations. In this way, considering the forcing for the Σ_α hierarchy from [1] and [2], we can prove the next generalization of Theorem 4.11.

If α is a constructive ordinal, $X \subseteq \mathbb{N}$, then by $X^{(\alpha)}$ we shall denote the α-th jump of X, see [4].

4.13. Theorem. *Let C and A be total sets. There exists a total set B such that $C \leq_T B$, $B \not\leq_{he} C$ and for all $X \subseteq \mathbb{N}$:*

(1) *For every constructive ordinal α, $X \leq_T A^{(\alpha)}$ & $X \leq_T B^{(\alpha)} \Rightarrow X \leq_T C^{(\alpha)}$;*

(2) *For every constructive ordinal α, if X is r. e. in $A^{(\alpha)}$ and X is r. e. in $B^{(\alpha)}$, then X is r. e. in $C^{(\alpha)}$;*

(3) $X \leq_h A \ \& \ X \leq_h B \Rightarrow X \leq_h C$;

(4) $X \leq_{he} A \ \& \ X \leq_{he} B \Rightarrow X \leq_{he} C$.

Other applications of the method will be presented in the forthcoming [7].

1. A s h, C., J. K n i g h t, M. M a n a s s e, T. S l a m a n. Generic copies of countable structures. — Ann. Pure Appl. Logic, **42**, 1989, 195–205.
2. C h i s h o l m, J. Effective model theory vs. recursive model theory. — J. Symbolic Logic, **55**, 1990, 1168–1191.
3. M c E v o y, K., S. B. C o o p e r. On minimal pairs of enumeration degrees. — J. Symbolic Logic, **50**, 1985, 983–1001.
4. R o g e r s, H. Theory of recursive functions and effective computability. McGraw-Hill Book Company, N. Y., 1967.
5. S a n c h i s, L. E. Hyperenumeration reducibility. — Notre Dame J. Formal Logic, **19**, 1978, 405–415.
6. S a n c h i s, L. E. Reducibilities in two models of combinatory logic. — J. Symbolic Logic, **44**, 1979, 221–233.
7. S o s k o v, I. N. Positive reducibilities on abstract structures (in preparation).
8. S o s k o v, I. N. Intrinsically Π^1_1 relations. — Mathematical Logic Quarterly, **42**, 1996, 109–126.

Received on 15.09.1996

Department of Mathematics and Computer Science
Sofia University
Blvd. "James Bourchier" 5
1164 Sofia, Bulgaria
E-mail address: soskov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

A SIMPLE PROOF OF A COINCIDENCE THEOREM OF RUBINSTEIN – WALSH AND GENERALIZATIONS*

PAVEL G. TODOROV

We give a simple proof of the Rubinstein – Walsh coincidence theorem that the classes of functions (1) and (2) can be represented in forms (4) and (5), respectively. We prove also that the more general classes of functions (8) and (9) can be represented in forms (4) and (5), respectively.

Keywords: coincidence theorem, subordination, rational functions, meromorphic functions.
Mathematics Subject Classification: 26C15, 30C45.

1. Let $R_1(D)$ and $R_2(D)$ denote the classes of rational functions

$$f(z) = \sum_{k=1}^{n} \frac{A_k}{z - a_k} \in R_1(D) \tag{1}$$

and

$$\varphi(z) := f\left(\frac{1}{z}\right) = \sum_{k=1}^{n} \frac{zA_k}{1 - a_k z} \in R_2(D), \tag{2}$$

respectively, where

$$\sum_{k=1}^{n} A_k = 1, \quad A_k > 0, \quad |a_k| \leqq 1, \quad 1 \leqq k \leqq n, \quad n \geqq 1. \tag{3}$$

* Lecture presented at the Session, dedicated to the centenary of the birth of Nikola Obreshkoff.

In [1, Lemma 2(a)] Rubinstein and Walsh prove that the functions (1) and (2) of the classes $R_1(D)$ and $R_2(D)$ can be represented in the corresponding forms

$$f(z) = \frac{1}{z - \alpha(z)} \tag{4}$$

for $|z| > 1$, and

$$\varphi(z) = \frac{z}{1 - z\beta(z)}, \quad \beta(z) := \alpha\left(\frac{1}{z}\right), \tag{5}$$

for $|z| < 1$, where $\alpha(z)$ and $\beta(z)$ are analytic functions with $|\alpha(z)| \leqq 1$ and $|\beta(z)| \leqq 1$ for $|z| > 1$ and $|z| < 1$, respectively. First we shall give a simple proof of this theorem of Rubinstein and Walsh.

Proof. For convenience we shall examine the class $R_2(D)$ only. From (2) and (3) we obtain

$$\operatorname{Re}\frac{\varphi(z)}{z} - \frac{1}{2} = \frac{1}{2}\sum_{k=1}^{n} A_k \operatorname{Re}\frac{1 + a_k z}{1 - a_k z} = \frac{1}{2}\sum_{k=1}^{n} A_k \frac{1 - |a_k z|^2}{|1 - a_k z|^2} > 0, \quad |z| < 1. \tag{6}$$

The inequality (6) shows that the function $\varphi(z)/z$ is subordinate to the function $1/(1-z)$ in $|z| < 1$, i. e.

$$\frac{\varphi(z)}{z} \prec \frac{1}{1 - z}, \quad |z| < 1. \tag{7}$$

According to the subordination (7) there exists an analytic function $\beta(z)$ in $|z| < 1$ satisfying $|\beta(z)| \leqq 1$, for which the representation (5) holds. If in (5) we replace z by $1/z$, we obtain (4).

This completes the proof.

2. Let M_1 and M_2 denote the more general classes of meromorphic functions with representations (4) and (5), respectively. In [2] we introduced the classes $S_1(D)$ and $S_2(D)$ of analytic functions

$$f(z) = \iint_D \frac{d\mu(\zeta)}{z - \zeta} \in S_1(D), \quad |z| > 1, \tag{8}$$

and

$$\varphi(z) := f\left(\frac{1}{z}\right) = \iint_D \frac{z\, d\mu(\zeta)}{1 - z\zeta} \in S_2(D), \quad |z| < 1, \tag{9}$$

respectively, where $D := \{\zeta \mid |\zeta| \leqq 1\}$ and $\mu(\zeta)$ is a unit mass measure on D, i. e.

$$\iint_D d\mu(\zeta) = 1, \quad d\mu \geqq 0. \tag{10}$$

If in (8) and (9) the unit mass is concentrated at n points of D, then, having in mind (10), we obtain sets $R_1(D)$ and $R_2(D)$ of rational functions (1) and (2) with the conditions (3), respectively. In the end of our paper [2] we put the problem

whether the classes $S_1(D)$ and $S_2(D)$ are corresponding subclasses of the classes M_1 and M_2 or not. Now we shall solve affirmatively this problem.

Theorem. *The classes $S_1(D)$ and $S_2(D)$ of functions* (8) *and* (9) *are corresponding subclasses of the classes M_1 and M_2 of functions* (4) *and* (5).

Proof. For convenience we shall examine the class $S_2(D)$ only. From (9) and (10) we obtain analogously

$$\operatorname{Re}\frac{\varphi(z)}{z}-\frac{1}{2}=\frac{1}{2}\iint\limits_{D}\frac{1-|z\zeta|^2}{|1-z\zeta|^2}\,d\mu(\zeta)>0,\quad |z|<1. \tag{11}$$

From (11) we obtain successively the subordination (7) and the representation (5) for the functions $\varphi(z)$ determined by (9) and (10). By replacing z by $1/z$ in (5), we obtain the representation (4) for the functions $f(z)$ determined by (8) and (10).

This completes the proof of the theorem.

Remark. If in (8) and (9) the unit mass is distributed on the circle C, $|\zeta|=1$, then, having in mind (10), we obtain the sets $S_1(C)$ and $S_2(C)$ of Schwarz analytic functions

$$f(z)=\int\limits_0^{2\pi}\frac{d\mu(t)}{z-e^{it}}\in S_1(C),\quad |z|>1,$$

and

$$\varphi(z):=f\left(\frac{1}{z}\right)=\int\limits_0^{2\pi}\frac{z\,d\mu(t)}{1-ze^{it}}\in S_2(C),\quad |z|<1,$$

respectively, where $\mu(t)$ is a probability measure on $[0,2\pi]$.

If in (8) and (9) the unit mass is distributed on the segment $[-1,1]$, then, having in mind (10), we obtain the sets N_1 and N_2 of Nevanlinna analytic functions

$$f(z)=\int\limits_{-1}^{1}\frac{d\mu(t)}{z-t}\in N_1,\quad |z|>1,$$

and

$$\varphi(z):=f\left(\frac{1}{z}\right)=\int\limits_{-1}^{1}\frac{z\,d\mu(t)}{1-zt}\in N_2,\quad |z|<1,$$

respectively, where $\mu(t)$ is a probability measure on $[-1,1]$.

According to the proved theorem the separate classes $S_{1,2}(C)$ and $N_{1,2}$ are corresponding subclasses of the classes $M_{1,2}$ as well.

REFERENCES

1. Rubinstein, Z., J. L. Walsh. Extension and some applications of the coincidence theorems. — Trans. Amer. Math. Soc., **146**, 1969, 413–427.
2. Todorov, P. G. Non-linear extremal problems for analytic two-dimensional Riemann-Stieltjes integrals. — Rendiconti del Circolo Matematico di Palermo, Serie II, T. XXXVII, 1988, 369–383.

Received on 03.05.1996

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bontchev str., Block 8
1113 Sofia, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

STUDY OF THE SCIENTIFIC WORK BY QUANTITATIVE METHODS: SOME RESULTS ON ACADEMICIAN NIKOLA OBRESHKOFF'S WORKS

VASSILIY TODOROV, MARA APOSTOLOVA, EMILIA BRANKOVA,
STEFKA ZLATEVA, VENETA TENEVA, DIMITAR KHRISTOV

Some results of studying the work of one of the most productive Bulgarian mathematician by quantitative methods are presented. The study is based on the data from the world-wide known review journals "Jahrbuch über die Fortschritte der Mathematik", "Zentralblatt für Mathematik und ihre Grenzgebiete" and others, representing most accurately the world scientific information flow, structuring it by domains of science and their areas. Graphically are shown: distribution of Obreshkoff's works over domains of mathematics according to divisions of mentioned review journals, distribution of scientific activity over years, domains of mathematics and their areas, etc.

This study is based on the so-called Reference Database (RDB) allowing flexible retrieving, systematizing, aggregation and generalizing data.

1. INTRODUCTION

Academician Nikola Obreshkoff is a Bulgarian scientist known not only to the Bulgarian mathematicians. He is respected by the whole Bulgarian scientific community for his over 40 years long scientific and publication activity. The goal of the present paper is to estimate by quantitative means his interference with the international scientific community.

Some results of using quantitative methods to explore his publication activity are presented in the paper. The notions of relevance criterion and the so-called Reference Database (RDB) are introduced. The data in the RDB on N. Obreshkoff are compared to the known bibliographies of his works. These bibliographies

are not used as sources to build a RDB, because they are lists of works, ordered chronologically or alphabetically. They are not organized according to the domains of scientific fields the scientist works in. No matter how complete they are, they do not give an adequate image of the interaction between the scientist and the international scientific community. This characteristic feature is the main reason to study the publication activity by RDB organized according to some classification of scientific domains.

2. WORLD-WIDE FLOW OF SCIENTIFIC INFORMATION AND RELEVANCE CRITERION

The mentioned interaction between scientists gives the so-called *world-wide flow of scientific information* built by an immense quantity of scientific works in different fields of science, published in numerous scientific journals, proceedings of conferences and workshops, monographs and so on. To manage that flow, the scientific community created the powerful tool of auxiliary reference editions — review (abstract) journals, bibliographies etc.

The consideration of the participation of a given scientific work in this information flow provides a useful bibliometric criterion — whether the paper has been or not reviewed in the world-widely known abstract journals. The use of that criterion when exploring the scientific work mirrors the publishing activity of given scientists and the dynamics of their scientific interests as the international scientific community looks at them.

Thus the idea is arisen of using the Reference Databases with published scientific works of one or more scientists — a computer database keeping data extracted from scientific reviews published in the abstract journals. Such a database can be explored by computer and quantitative tools from different points of view. This approach makes it possible to find some interesting and sometimes unexpected points in the entire work of a given scientist. The authors of the present paper are developing similar RDB, fulfilling the project "A quantitative study of the scientific production of lecturers of the Sofia University from 1889 to 1950"[1]. This project continues the research of the authors published in [5].

In this study the selection of sources is done following the above mentioned criterion: published works are taken into consideration only if they are reviewed in world-widely known abstract journals. These journals assign the reviews to sections in accordance with the domains of different fields of science. This is a good reason to use such journals for purposes of building RDB.

In the field of Mathematics the following journals were selected to build a RDB: *Jahrbuch über die Fortschritte der Mathematik* (Fortsch. d. Math.), *Zentralblatt für Mathematik und ihre Grenzgebiete* (Zbl. Math.), *Mathematical Reviews* and *Referativny Zhurnal*. The first one was founded in 1868 and was issued regularly until 1938[2], the second was founded in 1931, the third — in 1940, and the last — in 1953.

[1] Contract No 97/1996 of the Sofia University Scientific Research Fund.

[2] It was stopped several years later.

3. USING RDB TO PROCESS THE DATA ON N. OBRESHKOFF

The data in RDB concerning the works before 1939 are extracted from two abstract journals: *Jahrbuch über die Fortschritte der Mathematik* and *Zentralblatt für Mathematik und ihre Grenzgebiete*, and concerning the works after 1939 — from three abstract journals: *Zentralblatt für Mathematik und ihre Grenzgebiete*, *Mathematical Reviews* and *Referativny Zhournal*. The search in the journals was conducted for a period starting several years before 1920 (the year of N. Obreshkoff's entrance in the lecturer community of Sofia University) and continuing up to 1970, Vol. 178 of *Zbl. Math.* The assignment of entries to the sections and subsections before 1939 is made according to these of the *Fortsch. d. Math.*, and concerning the works after 1939 — according only to the sections and subsections of the *Zbl. Math.* The way of assignment is changed because the issuing of the first journal is suspended after 1939. Conforming all RDB to the classification before 1939 is useless. Thus, there is a boundary dividing the entire work of N. Obreshkoff into two periods: the first one from 1920 till 1939 (44% of the whole duration) and the second one from 1940 till 1963. For this reason works, for instance, belonging to the domain of Analysis, may have entries in section II (if the work is published before 1939) or in section V (if the work is published after 1939) in the RDB.

The creation of the RDB on N. Obreshkoff's work is based on a modification of a first variant of RDB on lecturers in the Faculty of Mathematics and Physics, built by the authors. This makes the investigation much easier.

4. RESULTS

A. A QUANTITATIVE INFORMATION ON SCIENTIFIC ACTIVITY IN THE PERIOD 1920–1939

The RDB has 99 entries for this period, assigned to the following domains in the field of Mathematics: I. Arithmetics and Algebra (21 reviewed works); II. Analysis (76 reviewed works); III. Geometry (2 reviewed works).

Fig. 1 shows the publication activity (the number of all works from 1920 to 1939) distributed over different domains. The domains I and III contain entries assigned to one area in each domain. The most of entries are in the domain of Analysis, assigned to several areas. Fig. 2 shows the distribution of the works over areas. It allows ranking the activity of N. Obreshkoff in this period. Thus, his scientific interests are oriented in the first place to the areas of *Infinite Number Sequences Theory* and *General Theory of Real Functions* (50% of all works). Near 31% of them are in the areas of *General Theory of Functions with Complex Arguments* and *Functions of Complex Variables*.

The scientific activity is often represented by the number of published works per year. The distribution of works in different domains per year is given on Fig. 3 representing the dynamics of scientific interests. Being concentrated in the domain of Analysis, the number of works varies — there is an alternation of decreasing and

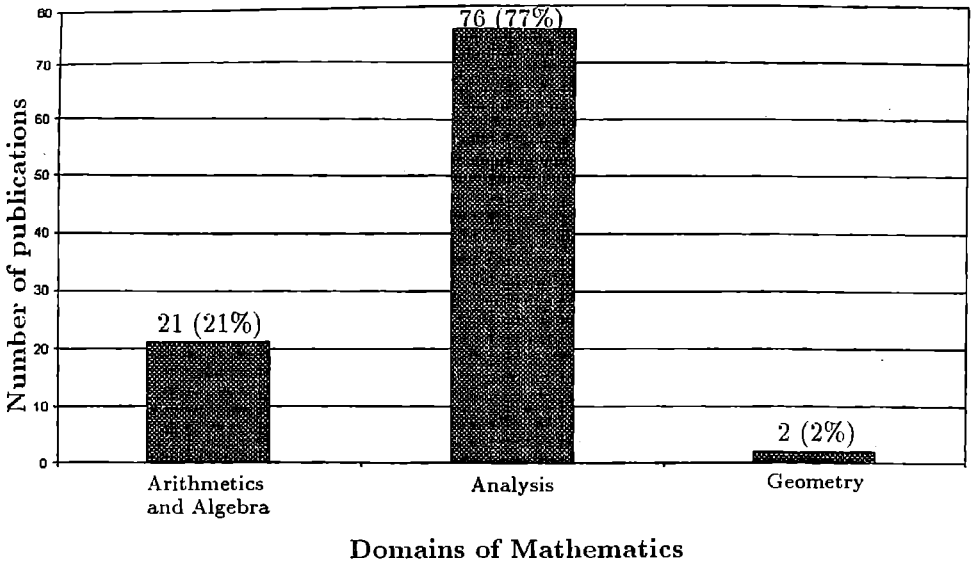


Fig. 1. Activity distribution (before 1939)

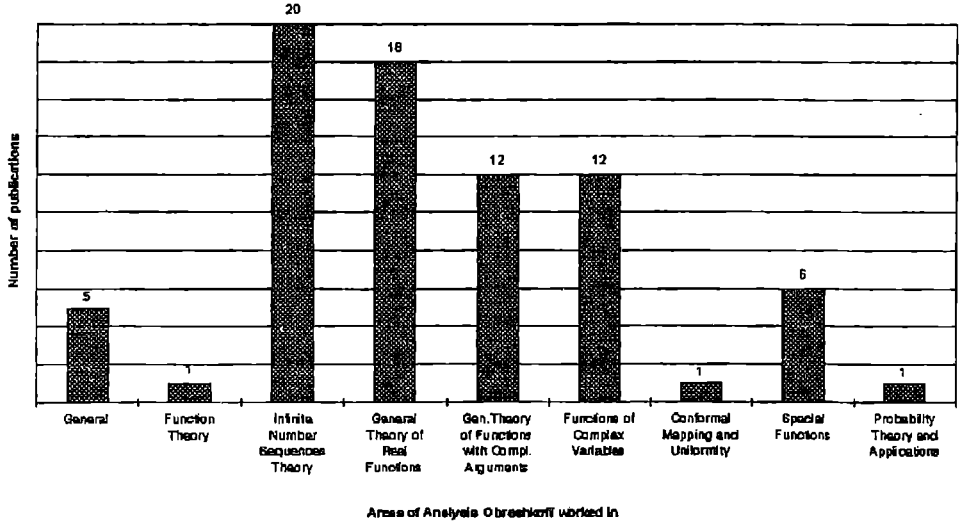


Fig. 2. Activity over areas of analysis (before 1939)

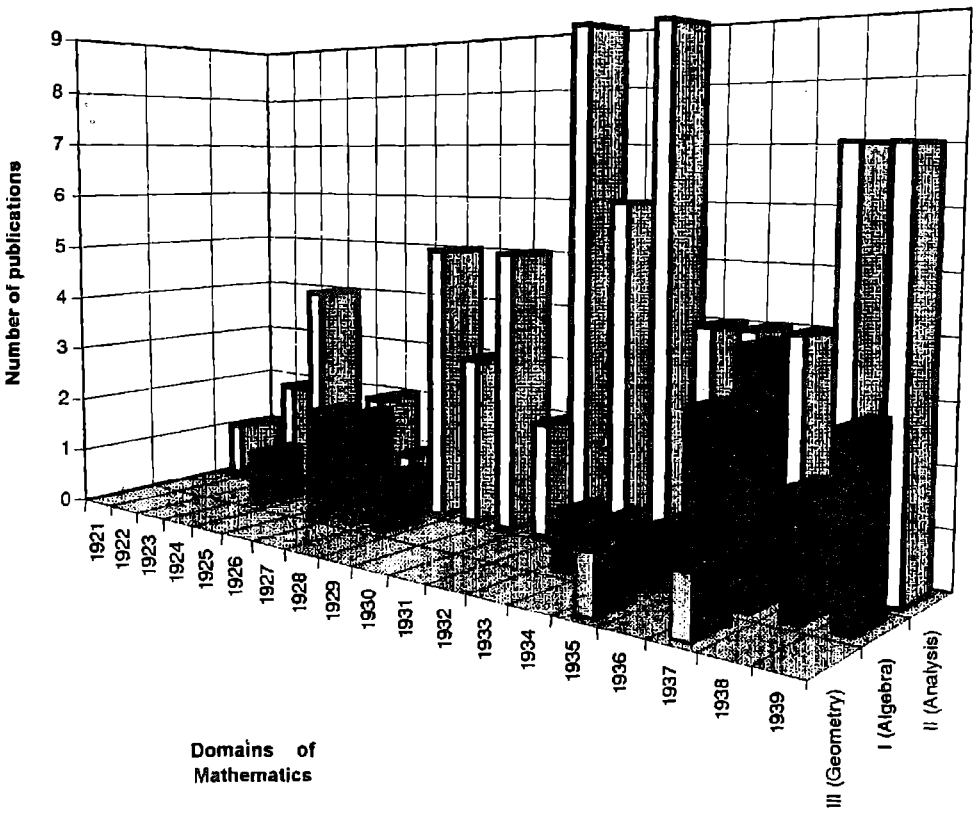


Fig. 3. Distribution of works over domains of mathematics per year (before 1939)

increasing in the activity in this domain; when the activity in the area of Analysis decreases, this one in the area of Algebra increases (a contre-tendence).

B. A QUANTITATIVE INFORMATION ON SCIENTIFIC ACTIVITY IN THE PERIOD 1940–1963

The RDB has 93 entries for this period assigned to the following domains in the field of Mathematics: IV. Algebra and Number Theory (34 reviewed works); V. Analysis (57 reviewed works); VI. Geometry (1 reviewed work), VII. Probability Theory. Statistics. Applications (4 reviewed works).

Fig. 4 shows the publication activity distribution over domains of Mathematics. It confirms the conclusion about concentration of interests in the domains of Algebra and Analysis.

Fig. 5 illustrates the activity over areas of analysis after 1939. In this period the classification is different compared with that of the first period. Nevertheless, Fig. 5 shows that the biggest part of published works is in the areas of *Real Function*

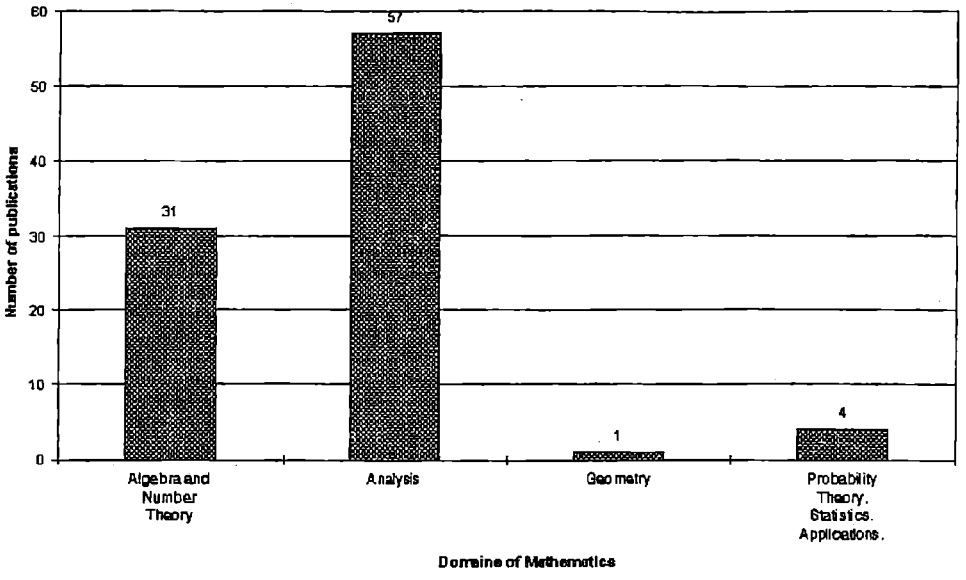


Fig. 4. Activity distribution (after 1939)

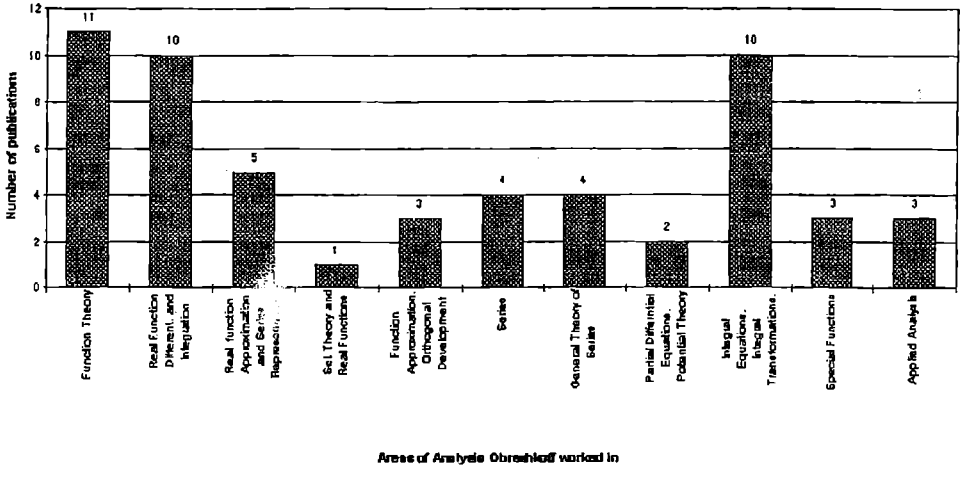


Fig. 5. Activity over areas of analysis (after 1939)

Differentiation and Integration and *Integral Equations, Integral Transformations.* There is a work not related to any area of Analysis, according to the subsections of *Zbl. Math.*, so the sum of the numbers in different areas is 56.

On Fig. 6 "Distribution of works over domains of Mathematics per year (after 1939)" the dynamics of N. Obreshkoff's works is shown. With concentration in the

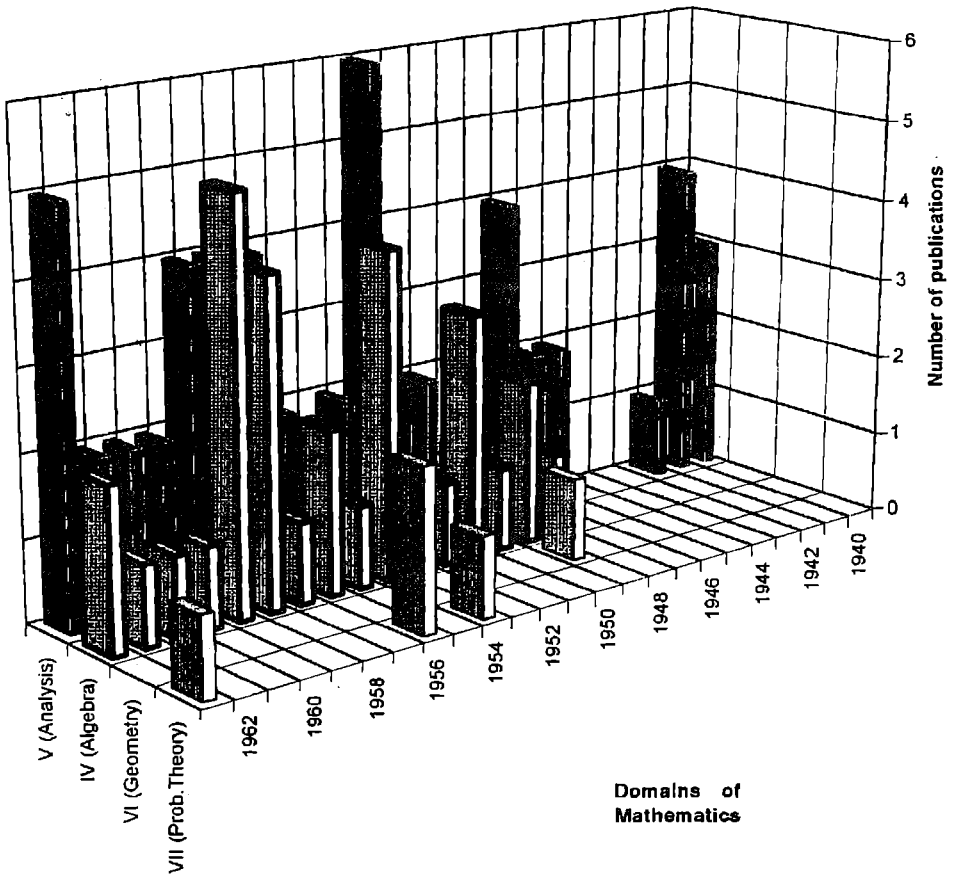


Fig. 6. Distribution of works over domains of mathematics per year (after 1939)

domains of Analysis and Algebra, similar to the first period, some equalising of tendencies near 1963 is observed. The number of works in the domain of Analysis prevails over the works in the area of Algebra near 1940. An interruption in the activity between 1943–1945 can be explained by the difficulties in publishing because of the World War II.

C. A GENERALIZED QUANTITATIVE INFORMATION ON THE ENTIRE SCIENTIFIC WORK IN THE PERIOD 1920–1963

The distribution of published works over the age of the scientist is given on Fig. 7. There is a period of extremely high activity starting in 1932 (when N. Obreshkoff was 36 years old) to 1939. The end of this period coincides with the beginning of the World War II. Here 67 published and reviewed works can be seen or 35% of all published and reviewed works. During this 7 year long period there are two absolutely maximal values of the activity (in 1934 and 1939). The second

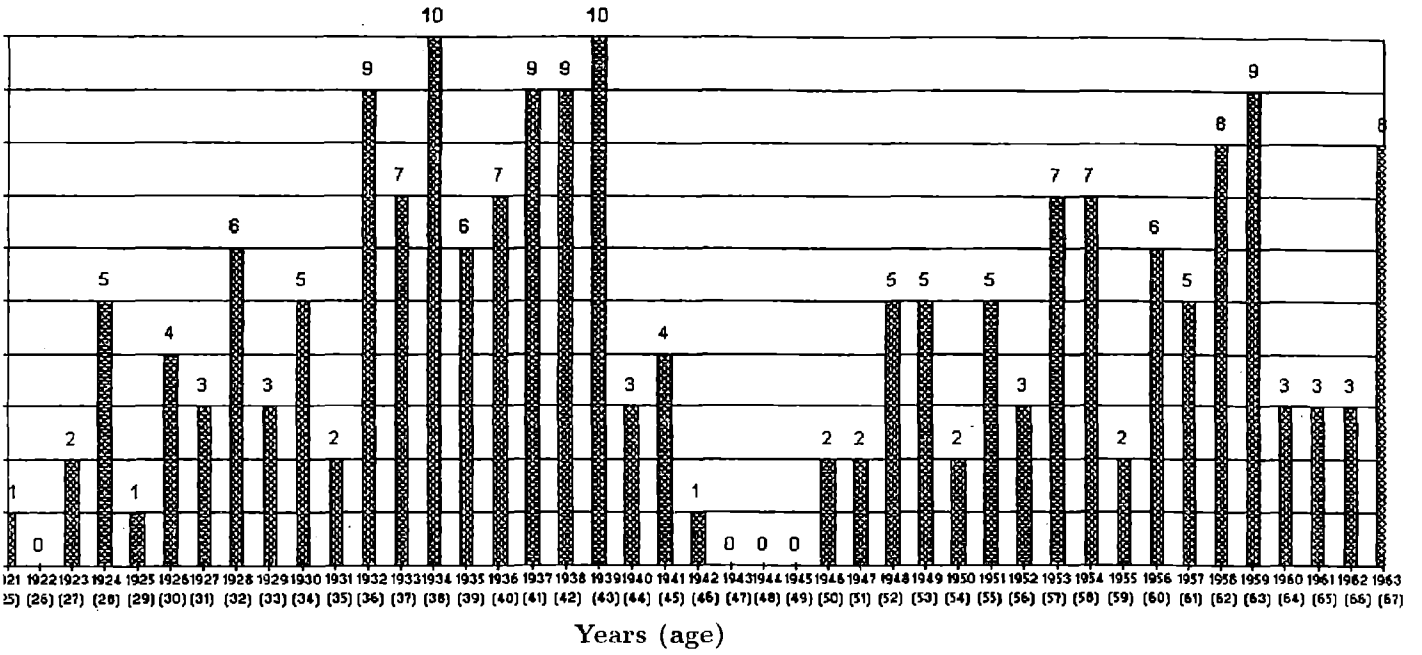


Fig. 7. Distribution of works of Nikola Obreshkoff

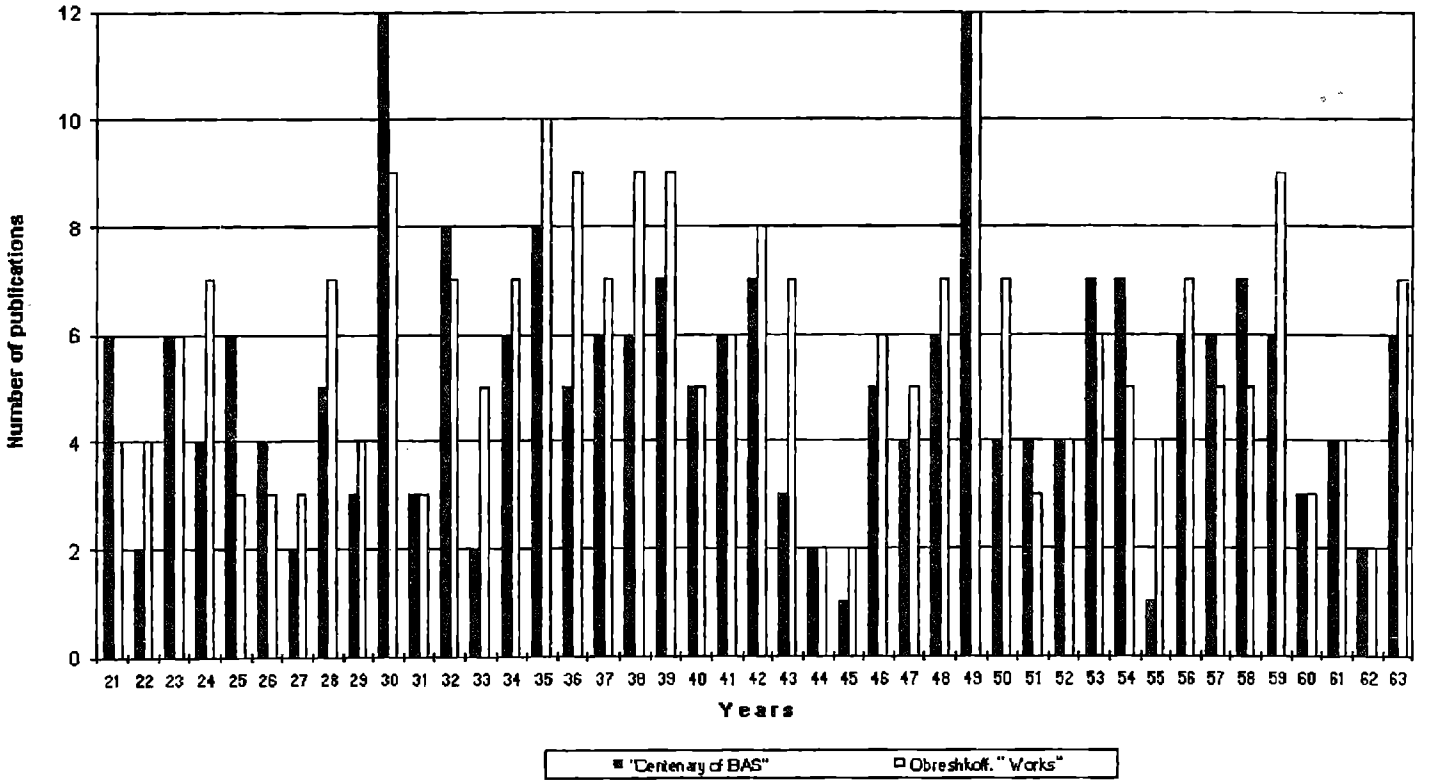


Fig. 8. Published works of Nikola Obreshkoff

maximum hints a new period of increase in the work of N. Obreshkoff, stopped by the beginning of the war.

It is interesting to compare our data with the known bibliographies of N. Obreskoff's works. The most complete one has 247 entries [6]. The bibliography in [4] has 219 entries. The bibliography in [3] includes works from 1940 to 1963. All of them were compiled after N. Obreshkoff's death in 1963. Two previous bibliographies are given in the first Almanacs of Sofia University the first one in 1929 [1] and the second one in 1940 [2]. They were compiled by Nikola Obreshkoff himself. The bibliography of 1929 includes entries missing in the later bibliographies, the one of 1940 is selective and its worth is Obreshkoff's own classification of works into groups of "principal works", "other works" and "diverse". Fig. 8 shows the distribution of published works over the years according to the biggest bibliographies [4, 6] which include not only reviewed works. The noticeable difference in 1930 can be explained by the fact that the Annuaire of the Sofia University was not reviewed before 1930. Another difference in 1949 can be explained by the difficulties in the cultural relations in Europe in the end of the World War II and after it.

Each domain of mathematics has two corresponding sections in this implementation of RDB. For this reason, in order to retrieve a quantitative information relative to the entire period from 1920 to 1963, the data are grouped into four domains: **A**. Algebra and Number Theory; **B**. Analysis; **C**. Geometry; **D**. Probability Theory. Statistics. Applications. The distribution of works over these domains is presented on Fig. 9.

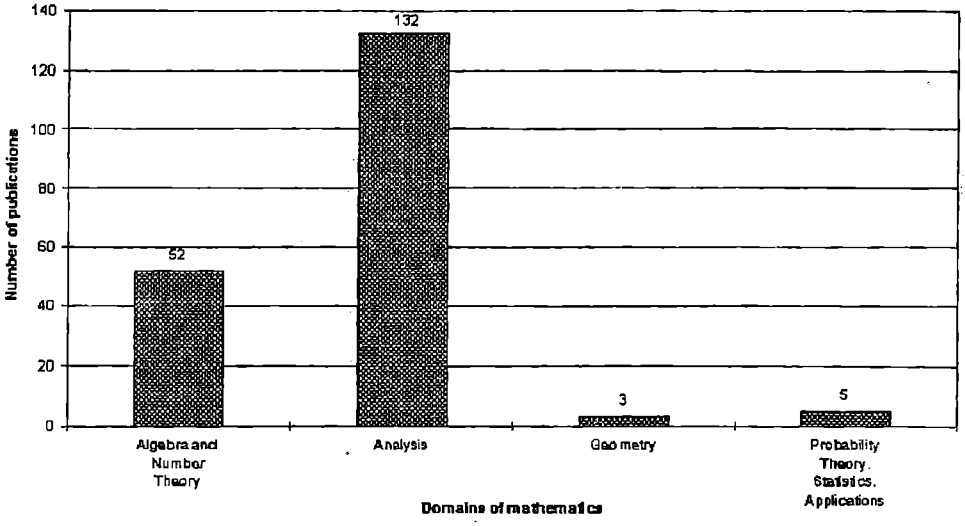


Fig. 9. Distribution of works over domains of mathematics

The reviewed works are published in nearly 50 journals, 3 monographs, 6 textbooks. Most of the papers are published in: Comptes Rendues Acad. Sci., Paris — 36 papers, Annuaire Univ. Sofia, Fac. Phys.-Math., Livre 1 — 25 papers, Comptes

Rendues Acad. Sci. Bulgare — 24 papers, Izvestia MI BAN — 14 papers, Jahresbericht D.M.V. — 11 papers.

5. CONCLUSIONS

The results presented in the paper allow to make: (i) deductions about the publication activity of N. Obreshkoff, and (ii) quantitative evaluation of dynamics of his scientific interests. Specific features of the activity like the contra-tendencies in its alternation are demonstrated.

The predominant orientation of interests to the Analysis and Algebra, shown graphically, can be compared to Obreshkoff's own view on his principal works [2]. There are 22 works, 15 in the area of Analysis (over 2/3) and only 7 in the area of Algebra. Of these principal works 15 are reviewed: 3 on Algebra (20%) and 12 (80%) on Analysis.

The results obtained show that the application of RDB was useful in exploring the work of the scientist. The data on scientific publication activity were considered according to different points of view. They were represented in different ways, and numeric evaluation, dynamics and distributions were obtained. The method of RDB is outlined as a necessary foundation in research on a scientist's publication and other activity, on its significance for evaluating the development of the corresponding scientific domain in Bulgaria and comparing it with the general tendencies in the development of the science in the world.

Last but not least, the RDB allows to explore the abstract journals themselves — their scope, degree of discordance in their classification schemes etc. This is an important area in research, based on the use of abstract journals.

REFERENCES

1. Almanac of Sofia University (1888-1928), Printing house "Khudozhnik", Sofia, 1929 (in Bulgarian).
2. Almanac of Sofia University "St. Kliment Ohridski", Second ed., Court printing house (Phototype edition, Publishing House of Sofia University), Sofia, 1940 (in Bulgarian).
3. Almanac of Sofia University (1939–1988), Publishing House of Sofia University, 1995 (in Bulgarian).
4. Centenary of Bulgarian Academy of Sciences (1869–1969), Vol. 1, Publishing House of Bulgarian Academy of Sciences, Sofia (in Bulgarian).
5. Khristov, D. et als. New Information on the History of Sofia University "St. Kliment Ohridski" undertaken by Quantitative Methods and Computer. 1888–1939., Publishing House of Sofia University, Sofia, 1990 (in Bulgarian).
6. Obreshkoff, N. Works. Vol. 1, Publishing House of Bulgarian Academy of Sciences, Sofia, 1977 (in Bulgarian).
7. Obreshkoff, N. Works. Vol. 2, Publishing House of Bulgarian Academy of Sciences, Sofia, 1981 (in Bulgarian).

Received on 04.07.1996

E-mail address: wassiliy@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

A FIRST-ORDER IN THICKNESS MODEL FOR FLEXURAL DEFORMATIONS OF GEOMETRICALLY NON-LINEAR SHELLS

CHRISTO I. CHRISTOV

The shallow shells, characterized by deflections of the order of unity, small deformations and still smaller curvatures, have most thoroughly been studied in the literature. However, the momentum terms, due to which the shell differs essentially from a membrane, are not negligible only for the short-wave-length deformations, when the deflections are small, the deformations — of the order of unity and the curvatures — of the order of the inverse of the small parameter. In order to treat consistently the case of momentum supporting shells, the formulas for covariant differentiation in the shell space are revisited. It is shown that the geometrical non-linearity contributes terms of the same order of magnitude as the momentum stresses. For the flexural deformations an equation of Boussinesq type is derived containing fourth-order dispersion and cubic non-linearity.

Keywords: shells, geometrical non-linearity, flexural deformations.
1991 Mathematics Subject Classification: 73K15.

1. INTRODUCTION

Since the turning of the century and especially in the late forties the theory of thin shells attracted much attention and many papers were devoted to its mechanical and mathematical aspects. Yet, it is far from completion. It goes beyond the framework of the present paper to give the historical account and the perspective of the numerous shell theories. We generally accept the attitudes of the comprehensive review [9] and the monographs [6, 8, 10] in assessing the vast body of the existing literature.

The theoretical approaches for modelling shells fall generally into two main groups. To the first belong the theories in which the governing equations are

derived as averaged properties of a very thin curved 2D elastic layer in the 3D space. The second approach originates in [14, 5] and consists in direct application of the mechanical laws to the 2D continuum representing the middle surface of the shell. The Cosserat concept was applied in [7]. For the problems arising in the asymptotic analysis of thin shells we refer the reader to the works of P. Ciarlet, E. Sanchez-Palencia and co-workers (see the recent works [4, 11] and the literature cited there).

When deriving the shell equations from the 3D elasticity, the deflections are assumed to be finite while the strains are small. This implies long wave length of the deformations, resulting in even smaller curvatures. This is the so-called "shallow shell" model. Strictly speaking, the shallow-shell approach is not generic for shells but it is rather adequate for membranes, because the momentum stresses that are supposed to make the difference between a shell and a membrane are proportional to the curvature of the deflections. Hence, in a consistent small-strains/smaller-curvatures approach, the moments are to be neglected to the first order of thickness unless the stiffness coefficient is extremely large. However, large values of the stiffness are very unlikely since the stiffness is proportional to bulk Young modulus and the square of the thickness, the latter being very small. Hence, the short length scale of the deformations is the case where the moment stresses are really important.

The difference between shells and membranes becomes really important when the strains are much larger than deflections, and curvatures — much larger than strains. It is clear that such a structure must be geometrically highly non-linear. We derive here a *consistent* first-order approximation in the shell thickness for the said case.

The assumptions of the present work are:

1. The thickness h of the shell is much smaller in comparison with the length scale L of the flexural deformations of the middle surface, i.e. $h \ll L$ or $\varepsilon \equiv h/L \ll 1$. No restrictions on L are imposed, e.g., $L \ll L_D$ is also an admissible case, where L_D is the length scale of the structure itself.
2. The thickness of the shell is constant within the adopted asymptotic order. Hence the derivatives of the thickness scaled by the thickness itself should not be large values, i.e. $\|h^{-1}(\nabla h)\| \approx O(1)$. The latter means that the length scale of changing the thickness is of order of magnitude larger than the length-scale of the deformations.
3. The loads, e.g. the normal pressure and the tractions on the shell faces, are compatible with the above assumptions, i.e. they possess the necessary asymptotic in order to secure 2D strain and stress states.
4. If the deformations created by the boundary conditions at the rim of the shell structure (the contour-line of the middle surface) are not compatible with (1) and (2), then only the portion of the shell is considered, which is far from the rim, i.e. the 3D effects of the said boundary conditions can be neglected.
5. For the sake of simplicity, no tractions are exerted on the shell faces.

It should also be mentioned that when the thickness of a shell is very small, then the contributions from the physical non-linearity of the material are negligible and geometry is the *only* source of non-linearity. For this reason, in the present

work we consider only the linear constitutive relations for elastic continuum (the so-called St-Venan–Kirchhoff materials [3]).

2. GEOMETRY OF THE SHELL SPACE

In this section we develop further the derivations of [12] and [6] incorporating the dependence on the transverse co-ordinate in the shell space. As it will turn out, this is essential, because after averaging some of the terms, neglected in the mentioned works, they become commensurable with those that had been left into the considerations.

Consider an N-dimensional Euclidean space and a structure immersed in it, defined as a thin layer of virtually constant thickness h (in the sense of requirement (1)). It is approximately equipartitioned (in the same sense) by the middle hypersurface of dimension $(N-1)$.

Assume that the middle surface is parameterized by the curvilinear co-ordinates ξ^α, $\alpha = 1, \ldots, N-1$. The N-th co-ordinate ξ^N is defined as the normal line to the particular point of the middle surface. As far as the shell does not intersect itself, the so defined set of curvilinear co-ordinates is not ambiguous. In addition, it is orthogonal and, within the adopted asymptotic order, it coincides with the material co-ordinates. When the shell thickness is not constant, then it is convenient to scale the normal co-ordinate by it, in order to transform the mathematical problem into one for which the shell faces are co-ordinate surfaces. Then the co-ordinate system is not strictly orthogonal but only to the order $O(\varepsilon^2)$, which is fully compatible with the attempted here theory of approximation $O(\varepsilon)$. We resort here to the case of equidistant surfaces of the shell and the words "equipartitioned by the middle surface" mean that the middle surface is drawn inside the shell, so that the condition $h_{\mathrm{lo}}(\xi^1, \ldots, \xi^{N-1}) = -h_{\mathrm{up}}(\xi^1, \ldots, \xi^{N-1})$, and hence $h \equiv h_{\mathrm{up}} - h_{\mathrm{lo}}$, always holds.

The curvilinear co-ordinates ξ^α, $\alpha = 1, \ldots, N-1$, are in fact material (Lagrangian) co-ordinates. They are connected to the geometrical Cartesian co-ordinates (originated somewhere in the ND-space) through the following functional dependences:

$$x^i = x^i(\xi^1, \ldots, \xi^N; t) \quad \text{for} \quad i = 1, \ldots, N\,, \tag{2.1}$$

where t stands for the time. Here and henceforth the Greek indices range from 1 to $N-1$ and serve to mark the variables in the shell. Italics are used for indices when the space quantities are concerned.

Let us assume for definiteness that the initial state of the shell is physically admissible (see, e.g., [13] for the definition). Then the initial state can be parameterized by the same transformation (2.1) but for the specific value of time $t = t_0$. Without loss of generality we set $t_0 = 0$.

The middle surface is characterized by the first and second fundamental forms

$$g_{\alpha\beta}(\xi^1, \ldots, \xi^N; t) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{N} \frac{\partial x^i}{\partial \xi^\alpha} \frac{\partial x^i}{\partial \xi^\beta}\,, \quad b_{\alpha\beta}(\xi^1, \ldots, \xi^N; t) \stackrel{\mathrm{def}}{=} -\sum_{i=1}^{N} \frac{\partial n^i}{\partial \xi^\alpha} \frac{\partial n^i}{\partial \xi^\beta}\,.$$

In the last formula n^i denote the Cartesian co-ordinates of the normal to the middle surface vector (say, $\boldsymbol{n}$). The outward normal is defined arbitrarily. When the co-ordinates are the lengths of the arcs, then the second fundamental form adopts the specially simple form $b_{\alpha\beta} = \nabla_\alpha \nabla_\beta \zeta$.

The orts of the curvilinear co-ordinate system are expressed as follows

$$\boldsymbol{g}_\alpha \stackrel{\text{def}}{=} \sum_{i=1}^{N} \frac{\partial x^i}{\partial \xi^\alpha} \boldsymbol{e}_\alpha ,$$

where $\boldsymbol{e}_\alpha$ are the orts of the Cartesian co-ordinate system. In order to avoid confusion, we do not use throughout the present work the convention of summation with respect to "dummy" indices when Cartesian co-ordinates are involved. In such a case we put explicit sign Σ. For the sake of completeness we also add the relation

$$\boldsymbol{g}_N \equiv \boldsymbol{n} ,$$

which is true by the definition of the normal co-ordinate. According to this definition the radius vector $\boldsymbol{r}$ of a point inside the ND-space enclosed in the shell can be expressed as

$$\boldsymbol{r} = \vec{\boldsymbol{r}} + s\boldsymbol{g}_N , \tag{2.2}$$

where $\vec{\boldsymbol{r}}$ is the radius-vector of the normal projection of the said point on the shell middle surface. Here we introduce the notation

$$s \equiv \xi^N h(\xi^1, \ldots, \xi^{N-1}) \tag{2.3}$$

as a measure of the length alongside the normal co-ordinate.

From Eqs. (2.2) and (2.3) one obtains for the fundamental tensor of the space enclosed in the shell (see [12, 6])

$$G_{\alpha\beta} = \left(\frac{\partial \boldsymbol{r}}{\partial \xi^\alpha} + s\frac{\partial \boldsymbol{n}}{\partial \xi^\alpha}\right) \cdot \left(\frac{\partial \boldsymbol{r}}{\partial \xi^\beta} + s\frac{\partial \boldsymbol{n}}{\partial \xi^\beta}\right) = \sum_{i=1}^{N} \left(\frac{\partial r^i}{\partial \xi^\alpha} + s\frac{\partial n^i}{\partial \xi^\alpha}\right)\left(\frac{\partial r^i}{\partial \xi^\beta} + s\frac{\partial n^i}{\partial \xi^\beta}\right)$$

$$\equiv g_{\alpha\beta}(\xi^1, \ldots, \xi^{N-1}) - 2sb_{\alpha\beta}(\xi^1, \ldots, \xi^{N-1}) + s^2 c_{\alpha\beta}(\xi^1, \ldots, \xi^{N-1}) , \tag{2.4}$$

$$G_{NN} = 1, \quad G_{\alpha N} = 0 . \tag{2.5}$$

Here $c_{\alpha\beta} = b_{\alpha\delta}b^\delta_\beta$ is the third fundamental form of the middle surface.

It is clear now that the fundamental tensor of the space filling the shell is defined both by the fundamental tensor of the middle hyper-surface (the first fundamental form) and by the tensor of curvature (the second fundamental form). For further convenience we cite here the formulas for the contravariant components of the fundamental tensor. Since our aim is a first order approximation with respect to thickness, it fully suffices to retain here only the terms up to $O(s^2)$.

Within the adopted order of approximation $o(s^2)$ the contravariant components of the fundamental tensor are given by

$$G^{\alpha\beta} = g^{\alpha\beta}(\xi^1, \ldots, \xi^{N-1}) + 2sb^{\alpha\beta}(\xi^1, \ldots, \xi^{N-1})$$

$$+3s^2 c^{\alpha\beta}(\xi^1, \ldots, \xi^{N-1}) + o(s^2) , \tag{2.6}$$

$$G^{NN} = 1, \quad G^{\alpha N} = 0 . \tag{2.7}$$

The proof of (2.7) is trivial and is a straightforward corollary of the definition of the matrix of contravariant components as an inverse matrix of the matrix of contravariant components. To prove (2.6), we simply multiply it by (2.4) to obtain

$$G_{\alpha\beta}G^{\alpha\gamma} = g_{\alpha\beta}g^{\alpha\gamma} + 2s(b_{\alpha\beta}g^{\alpha\gamma} - g_{\alpha\beta}b^{\alpha\gamma}) + s^2(c_{\alpha\beta}g^{\alpha\gamma} - 4b_{\alpha\beta}b^{\alpha\gamma} + 3g_{\alpha\beta}c^{\alpha\gamma}) + O(s^3)$$

$$= \delta^\gamma_\beta + 2s(b^\gamma_\beta - b^\gamma_\beta) + s^2(c^\gamma_\beta - 4c^\gamma_\beta + 3c^\gamma_\beta) + O(s^3) = \delta^\gamma_\beta + O(s^3)\,.$$

3. COVARIANT DIFFERENTIATION IN THE SHELL SPACE

This section uses extensively the results of [12] and [6], but it is not possible to omit it because not all of the necessary formulas are presented there. In addition, the terms proportional to s^2, which are essential for our derivations, are absent in the cited works. In order to make the present paper self-contained, on the one hand, and to fulfill the gaps in the cited works, on the other, we compile here the necessary formulas, deriving those that are not present in the literature.

The covariant derivatives of a vector and of a second-rank tensor are given by

$$A^n \Big\|_i = \frac{\partial A^n}{\partial \xi^i} + \Gamma^m_{ik}A^n\,, \quad A^{nm}\Big\|_i = \frac{\partial A^{mn}}{\partial \xi^i} + \Gamma^m_{ik}A^{kn} + \Gamma^n_{ik}A^{mk}\,. \tag{3.1}$$

The covariant Christofell symbol in N dimensions is given by

$$\Gamma_{ij,l} = \frac{1}{2}\left(\frac{\partial G_{jl}}{\partial x^i} + \frac{\partial G_{il}}{\partial x^j} - \frac{\partial G_{ij}}{\partial x^l}\right), \quad \Gamma^k_{ij} = G^{kl}\Gamma_{ij,l}\,.$$

The contravariant symbols are obtained from the covariant ones through the procedure of "elevation" ("contraction") of indices. It is easily shown now that a Christofell symbol is trivially equal to zero if it contains the index N at least in two positions, i.e.

$$\Gamma_{\alpha N,N} = \Gamma_{NN,\alpha} = \Gamma_{NN,N} = 0\,, \quad \Gamma^N_{\alpha N} = \Gamma^\alpha_{NN} = \Gamma^N_{NN} = 0 \quad \text{for} \quad \alpha = 1,\ldots,N-1\,.$$

Let us treat separately also the symbols containing the index N only in one position, namely:

$$\Gamma_{\alpha\beta,N} \equiv -\Gamma_{\beta N,\alpha} = -\frac{1}{2}\frac{\partial G_{\alpha\beta}}{\partial s} = b_{\alpha\beta} - sc_{\alpha\beta}\,.$$

Due to the specific properties of the fundamental tensor, namely, that $G^{Nj} = \delta^{Nj}$, one has

$$\Gamma^N_{\alpha\beta} = G^{Nj}\Gamma_{\alpha\beta,j} = \Gamma_{\alpha\beta,N} = b_{\alpha\beta} - sc_{\alpha\beta}\,.$$

Respectively,

$$\Gamma^\alpha_{\beta N} = \frac{1}{2}G^{\alpha\kappa}\frac{\partial G_{\beta\kappa}}{\partial s} = -(g^{\alpha\kappa} + 2sb^{\alpha\kappa} + 3s^2c^{\alpha\kappa})(b_{\beta\kappa} - sc_{\beta\kappa})$$

$$= -b^\alpha_\beta + sc^\alpha_\beta - s^2c^{\alpha\kappa}b_{\beta\kappa}\,.$$

Note that the last term is obtained after the following fairly obvious manipulation is applied $c^{\alpha\kappa}b_{\beta\kappa} = 3c^{\alpha\kappa}b_{\beta\kappa} - 2b^{\alpha\kappa}c_{\beta\kappa}$.

Finally, for the Christofell symbols which do not contain the index N, one derives

$$\Gamma_{\beta\gamma,\alpha} = [\beta\gamma,\alpha]^g - 2s[\beta\gamma,\alpha]^b + \frac{s^2}{2}[\beta\gamma,\alpha]^c\,, \tag{3.2}$$

where

$$\begin{aligned}
[\beta\gamma,\alpha]^g &\stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{\partial g_{\beta\alpha}}{\partial x^\gamma} + \frac{\partial g_{\gamma\alpha}}{\partial x^\beta} - \frac{\partial g_{\beta\gamma}}{\partial x^\alpha}\right),\\
[\beta\gamma,\alpha]^b &\stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{\partial b_{\beta\alpha}}{\partial x^\gamma} + \frac{\partial b_{\gamma\alpha}}{\partial x^\beta} - \frac{\partial b_{\beta\gamma}}{\partial x^\alpha}\right),\\
[\beta\gamma,\alpha]^c &\stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{\partial c_{\beta\alpha}}{\partial x^\gamma} + \frac{\partial c_{\gamma\alpha}}{\partial x^\beta} - \frac{\partial c_{\beta\gamma}}{\partial x^\alpha}\right)
\end{aligned} \tag{3.3}$$

are the connections generated by the tensors $g_{\alpha\beta}, b_{\alpha\beta}$ and $c_{\alpha\beta}$, respectively. One sees that due to the curvature of the middle surface the connections in the shell space are more complicated making its restriction to the $(N-1)$D-surface non–Riemannian. Note that the first term of the connections, namely ${}^g[\beta\gamma,\alpha]$, is nothing else but the Riemannian connection (ND-Christofell symbol) for the $(N-1)$-dimensional space of the middle surface.

The related contravariant Christofell symbol is expressed as usual

$$\Gamma^\alpha_{\beta\gamma} = G^{\alpha\kappa}\Gamma_{\beta\gamma,\kappa} = \left(g^{\alpha\kappa} + 2sb^{\alpha\kappa} + 3s^2c^{\alpha\kappa}\right)\left([\beta\gamma,\kappa]^g - 2s[\beta\gamma,\kappa]^b + \frac{s^2}{2}[\beta\gamma,\kappa]^c\right).$$

Then

$$\Gamma^\alpha_{\beta\gamma} = \left\{\begin{matrix}\alpha\\ \beta\gamma\end{matrix}\right\}^g + 2s\left\{\begin{matrix}\alpha\\ \beta\gamma\end{matrix}\right\}^b + s^2\left\{\begin{matrix}\alpha\\ \beta\gamma\end{matrix}\right\}^c,$$

$$\left\{\begin{matrix}\alpha\\ \beta\gamma\end{matrix}\right\}^g = g^{\alpha\kappa}[\beta\gamma,\kappa]^g\,, \quad \left\{\begin{matrix}\alpha\\ \beta\gamma\end{matrix}\right\}^b = b^{\alpha\kappa}[\beta\gamma,\kappa]^g - g^{\alpha\kappa}[\beta\gamma,\kappa]^b,$$

$$\left\{\begin{matrix}\alpha\\ \beta\gamma\end{matrix}\right\}^c = g^{\alpha\kappa}[\beta\gamma,\kappa]^c - 4g^{\alpha\kappa}[\beta\gamma,\kappa]^b + 3c^{\alpha\kappa}[\beta\gamma,\kappa]^g.$$

Now we are equipped to derive the expressions for the ND-covariant derivatives $\|_i$ for the space inside the shell. By definition we have

$$A^m\,\|_i = \frac{\partial A^m}{\partial\xi^i} + \Gamma^m_{in}A^n\,. \tag{3.4}$$

Let us also introduce the notation

$$A^\mu\,|_\alpha = \frac{\partial A^\mu}{\partial\xi^\alpha} + \left\{\begin{matrix}\mu\\ \alpha\nu\end{matrix}\right\}^g A^\nu\,, \tag{3.5}$$

which will be called "restriction of the covariant derivative." For $s = 0$ it is nothing else but the covariant derivative in the $(N-1)$D-space of the middle surface of the shell.

Since Eq. (3.5) is valid for the whole space inside the shell, it can only loosely be called "restriction of the covariant derivative". We shall return to this issue later on. For the time being it is enough to be noted that the only variables (3.5) that depend on the normal co-ordinate s are the components of the vector A^μ.

Combining Eqs. (3.5) and (3.4) and using the formulas for the Christofell symbols, one derives the following expressions for the covariant derivative $\|_i$:

$$A^\mu \Big\|_\alpha = A^\mu \Big|_\alpha + \left(2s\left\{ \begin{matrix} \mu \\ \nu\alpha \end{matrix} \right\}^b + s^2 \left\{ \begin{matrix} \mu \\ \nu\alpha \end{matrix} \right\}^c \right) A^\nu - (b^\mu_\alpha - sc^\mu_\alpha + s^2 c^{\mu\kappa} b_{\alpha\kappa}) A^N \, .$$

It is a generalization of the respective formula of Neuber because of the dependence on s of the components of the differentiated vector. Further on we have

$$A^N \Big\|_\alpha = A^N \Big|_\alpha + (b_{\nu\alpha} - sc_{\nu\alpha}) A^\nu = \frac{\partial A^N}{\partial \xi^\alpha} + (b_{\nu\alpha} - sc_{\nu\alpha}) A^\nu \, ,$$

because as far as the subspace of the middle surface is concerned, the component A^N behaves as a scalar, which means that

$$A^N \Big|_\alpha \equiv \frac{\partial A^N}{\partial \xi^\alpha} \, .$$

In the same manner we obtain

$$A^\alpha \Big\|_N = \frac{\partial A^\alpha}{\partial s} - (b^\alpha_\mu - sc^\alpha_\mu + s^2 c^{\alpha\kappa} b_{\nu\kappa}) A^\nu \quad \text{and} \quad A^N \Big\|_N = \frac{\partial A^N}{\partial s} \, .$$

Following the same line of reasoning, we obtain the formulas for the covariant differentiation of tensors:

$$A^{\alpha\beta} \Big\|_\gamma = A^{\alpha\beta} \Big|_\gamma + (2s[\nu\gamma, \alpha]^b + s^2 [\nu\gamma, \alpha]^c) A^{\nu\beta} + (2s[\nu\gamma, \beta]^b + s^2 [\nu\gamma, \beta]^c) A^{\alpha\nu}$$

$$- (b^\alpha_\gamma - sc^\alpha_\gamma + s^2 c^{\alpha\kappa} b_{\gamma\kappa}) A^{N\beta} - (b^\beta_\gamma - sc^\beta_\gamma + s^2 c^{\beta\kappa} b_{\gamma\kappa}) A^{\alpha N} \, ,$$

$$A^{\alpha N} \Big\|_\gamma = A^{\alpha N} \Big|_\gamma + (2s[\nu\gamma, \alpha]^b + s^2 [\nu\gamma, \alpha]^c) A^{\nu N}$$

$$+ (b_{\nu\gamma} - sc_{\nu\gamma}) A^{\alpha\nu} - (b^\alpha_\gamma - sc^\alpha_\gamma + s^2 c^{\alpha\kappa} b_{\gamma\kappa}) A^{NN} \, ,$$

$$A^{N\beta} \Big\|_\gamma = A^{N\beta} \Big|_\gamma + (2s[\nu\gamma, \beta]^b + s^2 [\nu\gamma, \beta]^c) A^{N\nu}$$

$$+ (b_{\nu\gamma} - sc_{\nu\gamma}) A^{\nu\beta} - (b^\beta_\gamma - sc^\beta_\gamma + s^2 c^{\beta\kappa} b_{\gamma\kappa}) A^{NN} \, ,$$

$$A^{NN} \Big\|_\gamma = A^{NN} \Big|_\gamma + (b_{\nu\gamma} - sc_{\nu\gamma}) A^{N\nu} + (b_{\nu\gamma} - sc_{\nu\gamma}) A^{\nu N} \, .$$

In the end we consider $A^{N\beta}$ and $A^{\alpha N}$, which are in fact components of a vector as far as differentiation in the middle surface of the shell is concerned:

$$A^{\alpha N} \Big\|_N = \frac{\partial A^{\alpha N}}{\partial s} - (b^\alpha_\nu - sc^\alpha_\nu + s^2 c^{\alpha\kappa} b_{\nu\kappa}) A^{\nu N} \, ,$$

$$A^{N\beta}\,\Big\|_N = \frac{\partial A^{N\beta}}{\partial s} - (b^\beta_\mu - sc^\beta_\mu + s^2 c^{\beta\kappa} b_{\nu\kappa}) A^{N\mu}\,, \quad A^{NN}\,\Big\|_N = \frac{\partial A^{NN}}{\partial s}\,.$$

Let us note again that our derivations are not restricted (as it is the case with [12] and [6]) to the middle surface but are valid for the entire shell space.

4. GOVERNING EQUATIONS IN CAUCHY FORM

We prefer to derive in the beginning the averaged Cauchy form and only after that to turn to constitutive relations, because even when considering stress balance, the role of geometrical non-linearity is conspicuous. The Cauchy form of the balance laws for a continuous media reads

$$\left[\,\rho_* a^j - P^{ij}\,\|_i - F^j\,\right] g_j = 0, \quad i,j = 1,\ldots,N\,, \tag{4.1}$$

where ρ_* is the ND-density of the elastic medium filling the shell; g_j are the above defined orts of the curvilinear co-ordinate system; P^{ij} are the components of stress tensor; a^j are the components of the acceleration vector and F^j — the components of the N-dimensional body forces. Respectively, $\|_i$ stands for the covariant derivative in $(N-1)$-dimensional space.

Upon substituting into Eq. (4.1) the above defined connection of $\|_i$ to the $(N-1)$D-covariant derivatives $|_\alpha$, the Cauchy law (4.1) is recast into a system for the "surface" (laminar) components and a scalar equation for the N-th component, namely

$$\rho_* a^\alpha - P^{\beta\alpha}\,\Big|_\beta = \frac{\partial P^{N\alpha}}{\partial s} - (b^\beta_\beta - sc^\beta_\beta + s^2 c^{\beta\kappa} b_{\beta\kappa}) P^{N\alpha}$$

$$- 2(b^\alpha_\nu - sc^\alpha_\nu + s^2 c^{\alpha\kappa} b_{\nu\kappa}) P^{N\nu} + 2\left(2s\left\{\begin{matrix}\alpha\\ \beta\nu\end{matrix}\right\}^b + s^2\left\{\begin{matrix}\alpha\\ \beta\nu\end{matrix}\right\}^c\right) P^{\nu\beta} + o(s^2)\,, \tag{4.2}$$

$$\rho_* a^N - P^{\beta N}\,\Big|_\beta = \frac{\partial P^{NN}}{\partial s} + (b_{\beta\nu} - sc_{\beta\nu}) P^{\beta\nu}$$

$$- (b^\beta_\beta - sc^\beta_\beta + s^2 c^{\beta\kappa} b_{\beta\kappa}) P^{NN} + \left(2s\left\{\begin{matrix}\beta\\ \beta\nu\end{matrix}\right\}^b + s^2\left\{\begin{matrix}\beta\\ \beta\nu\end{matrix}\right\}^c\right) P^{\nu N} + F^N + o(s^2)\,. \tag{4.3}$$

We simplify the above system by taking into account the main assumptions of the present derivations, namely that the shell is a thin layer $h \ll 1$ and that the length-scale of the deformations in the middle surface is $L \gg h$, then we have the small parameter $\varepsilon = h/L$. Dimensionless variables are introduced as follows:

$$s = hs'\,, \quad |_\alpha \simeq L^{-1}\,, \quad b_{\alpha\beta} = L^{-1} b'_{\alpha\beta}\,, \quad c_{\alpha\beta} = L^{-2} c'_{\alpha\beta}\,, \quad P_{ij} = \mu P'_{ij}\,,$$

$$\left\{\begin{matrix}\beta\\ \alpha\nu\end{matrix}\right\}^b = L^{-1}\left\{\begin{matrix}\beta\\ \alpha\nu\end{matrix}\right\}^b, \quad \left\{\begin{matrix}\beta\\ \alpha\nu\end{matrix}\right\}^c = L^{-2}\left\{\begin{matrix}\beta\\ \alpha\nu\end{matrix}\right\}^c,$$

$$t = \frac{L}{c\sqrt{\delta}}\,t'\,, \quad c = \sqrt{\frac{\mu}{\rho_*}} \quad \Rightarrow \quad a^\alpha = \delta\,\frac{c^2}{L}\,a'^\alpha\,, \quad a^N = \delta\,\frac{c^2}{L}\,a'^N\,; \tag{4.4}$$

here μ is the shear elastic modulus and c is the speed of shear waves. Note the special scaling for the time involving the square root of the parameter δ, which will

be identified later on. In a sense we consider motions of the shell that are of certain characteristic time. Omitting the primes without fear of confusion, the governing equations (4.2) and (4.3) read

$$\delta a^{\alpha} - P^{\beta\alpha}\Big|_{\beta} = \frac{1}{\varepsilon}\frac{\partial P^{N\alpha}}{\partial s} - (b_{\beta}^{\beta} - s\varepsilon c_{\beta}^{\beta} + s^2\varepsilon^2 c^{\beta\kappa} b_{\beta\kappa}) P^{N\alpha}$$

$$-2(b_{\nu}^{\alpha} - s\varepsilon c_{\nu}^{\alpha} + s^2\varepsilon^2 c^{\alpha\kappa} b_{\nu\kappa}) P^{N\nu} + 2\left(2s\varepsilon \left\{ \begin{matrix} \alpha \\ \beta\nu \end{matrix} \right\}^{b} + s^2\varepsilon^2 \left\{ \begin{matrix} \alpha \\ \beta\nu \end{matrix} \right\}^{c}\right) P^{\nu\beta} + o(\varepsilon^2)\,, \quad (4.5)$$

$$\delta a^{N} - P^{\beta N}\Big|_{\beta} = \frac{1}{\varepsilon}\frac{\partial P^{NN}}{\partial s} + (b_{\beta\nu} - \varepsilon s c_{\beta\nu}) P^{\beta\nu}$$

$$-(b_{\beta}^{\beta} - s\varepsilon c_{\beta}^{\beta} + s^2\varepsilon^2 c^{\beta\kappa} b_{\beta\kappa}) P^{NN} + \left(2s\varepsilon \left\{ \begin{matrix} \beta \\ \beta\nu \end{matrix} \right\}^{b} + s^2\varepsilon^2 \left\{ \begin{matrix} \beta \\ \beta\nu \end{matrix} \right\}^{c}\right) P^{\nu N} + o(\varepsilon^2)\,. \quad (4.6)$$

It is too early to make here assumptions about the relative asymptotic order of the different stress components. Yet one can compare the terms containing the same stress component and to neglect those which are of higher asymptotic order. Since we only consider here the flexural deformations, we can neglect the acceleration terms in the equations for the laminar components of motion. Thus we obtain

$$-P^{\beta\alpha}\Big|_{\beta} = \frac{1}{\varepsilon}\frac{\partial P^{N\alpha}}{\partial s}\,, \quad (4.7)$$

$$\delta a^{N} - P^{\beta N}\Big|_{\beta} = \frac{1}{\varepsilon}\frac{\partial P^{NN}}{\partial s} + (b_{\beta\nu} - \varepsilon s c_{\beta\nu}) P^{\beta\nu}\,. \quad (4.8)$$

The essential component of derivation of any kind of shell theory is the introduction of averaged across the shell variables, namely

$$\sigma^{\alpha\beta} \stackrel{\text{def}}{=} \int P^{\alpha\beta}\,\mathrm{d}s\,, \quad m^{\alpha\beta} \stackrel{\text{def}}{=} \int sP^{\alpha\beta}\,\mathrm{d}s\,, \quad q^{\alpha} \stackrel{\text{def}}{=} \int P^{N\alpha}\,\mathrm{d}s\,. \quad (4.9)$$

Integrating the asymptotically reduced equation (4.7), we get

$$\sigma^{\alpha\beta}\Big|_{\beta} = 0\,, \quad (4.10)$$

where it is acknowledged that there are no tractions on the shell faces. The last equation has an obvious solution

$$\sigma^{\alpha\beta} = \kappa_0 g^{\alpha\beta}\,, \quad (4.11)$$

which, depending on the sign of κ_0, corresponds to the case of uniform compression/dilation of the middle surface of the shell. Such a stress state is possible without motion in the middle surface. Henceforth we shall consider only the flexural deformations and the most complicated stress state in the middle surface will be given by Eq. (4.11).

Multiplying Eq. (4.7) by s, integrating and discarding the tractions on the faces, we get

$$\varepsilon m^{\alpha\beta}\Big|_{\beta} = q^{\alpha}\,. \quad (4.12)$$

Let us assume now that on the shell faces different normal pressures act with difference of order of $O(\varepsilon)$. Then

$$P^{NN}\Big|_{s=-\frac{1}{2}} = 0\,, \quad P^{NN}\Big|_{s=\frac{1}{2}} = \varepsilon V_g\,,$$

where εV_g stands for the pressure difference. Here it becomes clear that one can have effectively 2D stress and strain fields only when the normal pressure is of the above adopted order in the small parameter.

Integrating Eq. (4.8) with respect to s, taking into account the boundary conditions for P^{NN} and using Eq. (4.12), yields

$$\frac{\delta}{\varepsilon}\int a^N \,\mathrm{d}s = m^{\alpha\beta}\,|_\beta|_\alpha + \frac{\kappa_0}{\varepsilon} b_{\beta\nu} g^{\beta\nu} - c_{\beta\nu} m^{\beta\nu} + \frac{1}{2} V_g\,. \tag{4.13}$$

Obtaining the last equation has been the primary objective of the present paper, because it gives the opportunity to identify the geometrical non-linearity, namely the terms of type $c_{\beta\nu} m^{\beta\nu}$ containing the third fundamental form of the middle surface. Now it becomes clear that the spatial derivatives of the moment stresses are of the *same order* as the geometrical non-linearity. This is a new result and it is obtained due to the more consistent treatment of the covariant derivatives in the shell space in comparison with [12, 6].

5. CONSTITUTIVE RELATIONS. St-VENAN–KIRCHHOFF MATERIALS

We shall not dwell much on the constitutive relations for the shell. The main assumption is that for the very thin shells under consideration the material non-linearity is negligible and that the hypothesis of Kirchhoff–Love holds true. According to the latter, the laminar displacements u_α in the shell space are related to the $(N-1)$D-displacements $\tilde{u}_\alpha$ in the shell middle surface as follows:

$$u^\alpha = \tilde{u}^\alpha - \varepsilon s \nabla^\alpha \zeta\,. \tag{5.1}$$

Being consistent with the limiting case of flexural deformation, we neglect in what follows the lamin components $\tilde{u}_{\alpha\beta}$ of the displacement vector. Respectively, the transverse (flexura displacement and the acceleration, due to the latter, are given by

$$u^N = \zeta \quad \Longrightarrow \quad \int a^N \,\mathrm{d}s = \varepsilon \frac{\partial^2 \zeta}{\partial t^2}\,.$$

We consider an elastic material (called St-Venan–Kirchhoff material) whose constitutive relations are linear regardless to the presence or absence of geometrical non-linearity (see the thorough treatment of these materials in [2]). Without going into much detail one can derive the following linear constitutive relations for the averaged stresses and momenta in the middle surface:

$$m^{\alpha\beta} = -\bar{D} b^{\alpha\beta} \equiv -\bar{D} \nabla^\alpha \nabla^\beta \zeta\,, \tag{5.2}$$

where

$$\bar{D} = \frac{Dh}{\mu L^2} D$$

is the dimensionless stiffness coefficient, while D is the stiffness of shell. Alternatively, under the same assumptions the constitutive relation for the moment stresses can be postulated (see, [7]) and then the hypothesis of Kirchhoff–Love (5.1) is not necessary. Furthermore, the overbar will be omitted without fear of confusion.

Introducing Eq. (5.2) into Cauchy equations we get

$$\frac{\delta}{\varepsilon}\frac{\partial^2\zeta}{\partial t^2} = D\left[-\Delta\Delta\zeta + (\nabla_\beta\nabla_\delta\zeta)(\nabla^\beta\nabla_\mu\zeta)(\nabla^\mu\nabla^\delta\zeta)\right] + \frac{\kappa_0}{\varepsilon}\Delta\zeta + V_g\,, \qquad (5.3)$$

where $\Delta \equiv \nabla_\nu\nabla^\nu$, $\Delta\Delta \equiv \nabla_\nu\nabla^\nu(\nabla_\kappa\nabla^\kappa)$.

Now it is time to assess the length and time scales for which the momentum stresses are important, i.e. when the shell is not essentially a membrane. These scales are the ones for which the different coefficients in Eq. (5.3) are of the same order. For the sake of brevity, let us consider the case $V_g = 0$ when the normal load is absent. In fact, one can think that either the shell is a vast sheet, compressed at its rims, or a sphere subjected to normal pressure. In the second case, part of the membrane stress is balanced by V_g and one can subtract $V_g g_{\alpha,\beta}$ from the term $\kappa_0 b_{\alpha\beta}$. As a result the normal pressure drops off from the equation and its sole role is to create the uniform compression.

Thus the uniform membrane tension must be of order

$$|\kappa_0| = \frac{Dh^2}{\mu L^3} \qquad (5.4)$$

and the dimensionless time scale $\delta = |\kappa_0|$. Conversely, for a shell of given stiffness and shear modulus Eq. (5.4) defines the length scale of the "shell-type" deformations when the uniform compression/dilation κ_0 is selected. The governing equation then reads

$$\frac{\partial^2\zeta}{\partial t^2} = \left[-\Delta\Delta\zeta + (\nabla_\beta\nabla_\delta\zeta)(\nabla^\beta\nabla_\mu\zeta)(\nabla^\mu\nabla^\delta\zeta)\right] + \mathrm{sign}(\kappa_0)\Delta\zeta\,. \qquad (5.5)$$

One sees that Eq. (5.5) contains a very strong non-linearity — the cubic power of the curvature of the deformation. In this way it looks very much like the Boussinesq equation [1], being in fact a Boussinesq equation for the curvature $\Delta\zeta$, if the middle surface is subjected to uniform dilation $\kappa_0 > 0$. For the opposite case $\kappa_0 < 0$, when there is a uniform compression, it is more proper to be called *anti*-Boussinesq equation.

6. CONCLUSIONS

In the present paper a consistent asymptotic treatment of a 3D thin elastic layer is attempted for the purposes of derivation of shell theory. The main small parameter is the ratio between the thickness of the shell and the length scale of the deformation of the middle surface. No additional assumptions, such as "shallowness" of the flexural deformation, are implied. For the "steeper" deflections the geometrical non-linearity is identified and shown to be proportional to the cubic power of the curvature of the middle surface. The equation for flexural deformations turns out to be a Boussinesq-like equation.

ACKNOWLEDGEMENTS. This work was partially supported by the National Science Foundation of Bulgaria under Grant NZ-611/96.

REFERENCES

1. Boussinesq, J. V. Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. *J. de Math. Pures et Appl.*, **17**, Ser. 2, 1872, 55–108.
2. Ciarlet, P. Mathematical Elasticity. I: Three-Dimensional Elasticity. North-Holland, Amsterdam, 1988.
3. Ciarlet, P. Mathematical Elasticity. II: Lower-Dimensional Theories of Plates and Rodes. North-Holland, Amsterdam, 1988.
4. Ciarlet, P., V. Lods. Asymptotic Analysis of Linearly Elastic Shells: "Generalized Membrane Shells." *J. of Elasticity*, **43**, 1996, 147–188.
5. Chien , W. Z. The Intrinsic theory of thin shells and plates. *Quart. Appl. Math.*, **1**, 1943, 297–327; **2**, 1944, 43–59, 120–135.
6. Dikmen, M. Theory of Thin Elastic Shells. Pitman, Boston, 1982.
7. Ericksen, J. L., C. Truesdell. Exact theory of stress and strain in rods and shells. *Arch. Rat. Mech. Anal.*, 1, 1959, 295–323.
8. Goldenveizer, A. V. Theory of Elastic Thin Shells. Moscow, 1976, 2nd Edition (in Russian).
9. Koiter, W. T., J. B. Simmons. Foundations of Shell theory. In: *Proc. 13th Int. Congr. Theor. Appl. Mech., Moscow, 1972*, eds. E. Becker and G. K. Mikhailov, Springer, Berlin, 1972, 150–176.
10. Mathuna, O. Mechanics, Boundary Layers, and Functional Spaces. Dublin, 1989.
11. Miara, B., E. Sanchez–Palencia. Asymptotic Analysis of Linearly Elastic Shells. *Asymptotic Analysis*, **12**, 1996, 41–54.
12. Neuber, H. Allgemeine Schalentheorie. *ZAMM*, **29**(4), 1949, 97–108, 142–146.
13. Sedov, L. I. Mechanics of Continua. Nauka, Moskow, 1972 (in Russian).
14. Synge, J. L., W. Z. Chien. The intrinsic theory of elastic shells and plates. In: *Th. v. Kármán Anniv. Volume*, 1941, 103.

Received on July 25, 1996

National Institute of Meteorology and Hydrology
Bulgarian Academy of Sciences
BG-1184 Sofia, Bulgaria
e-mail: christo.christov@meteo.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

PERTURBATIONS IN A CHAMPAGNE BOTTLE

GEORGY GEORGIEV

The system describing the motion of a particle in a potential field shaped like the bottom of a champagne bottle (more precisely, an S^1 symmetric double well) for the KAM-theory conditions is studied. We show that the Kolmogorov's condition is fulfilled everywhere out of the bifurcation diagram of the energy-momentum map and we make researches for the condition of isoenergetical non-degeneracy.

Keywords: KAM-theory, Abelian Integrals, Kolmogorov's condition, Isoenergetical non-degeneracy.
Mathematics Subject Classification: 34D10, 58F07, 58F30, 70H05.

1. INTRODUCTION

The question of the integrability of Hamiltonian systems is one of most important problems of the classical mechanics (see [1]). Since the end of the last century it has been known that most of the Hamiltonian systems are not integrable. The main problem after this result is to study Hamiltonian systems which are close to integrable ones. The most powerful approach to non-integrable systems is the perturbation theory and especially the KAM-theory. Important for the KAM-theory are the conditions of non-degeneracy and isoenergetical non-degeneracy.

Before giving a brief account of KAM-theory, let us display the structure of the integrable Hamiltonian system (see Ch. 2 and [1] for details). The phase space of a general integrable Hamiltonian system with n degrees of freedom is foliated into invariant manifolds, the typical fiber being an n-dimensional torus on which the motion is quasiperiodic. As most of the motions of generic integrable systems are quasiperiodic, it is a logical question whether small perturbations can

destroy them. KAM-theory [1, 3] gives conditions for the integrable systems which ensure the survival of most of the invariant tori. One typical condition is that the frequency map should be a local diffeomorphism. For any integrable Hamiltonian system defined by a Hamiltonian H_0 one can introduce at least locally near a fixed torus canonical co-ordinates $I_1, \ldots, I_n$, $\varphi_1, \ldots, \varphi_n$ such that $I = (I_1, \ldots, I_n)$ maps a neighbourhood of the fixed torus into an open subset of $\mathbf{R}^n$ and $\varphi = (\varphi_1, \ldots, \varphi_n)$ are co-ordinates on any of the nearby tori. Moreover, the first integrals become functions only of $I_1, \ldots, I_n$. The theorem stated by Kolmogorov [3] maintains that in the perturbed system

$$H(I, \varphi) = H_0(I) + \varepsilon H_1(I, \varphi) \ ,$$

defined by a small Hamiltonian perturbation of H_0, most of the tori sustain the perturbation, provided that the Hesseian

$$\det \left(\frac{\partial^2 H_0}{\partial I^2} \right) \tag{1.1}$$

is not identically zero. The measure of the surviving tori decreases with the increase of both the perturbation and the measure of the set, where the above Hesseian is sufficiently close to zero.

In this paper we study the frequency map

$$I \to (\omega_1(I), \ldots, \omega_n(I)),$$

where

$$\omega_i(I) = \frac{\partial H_0}{\partial I_i} \ , \qquad i = 1, \ldots, n,$$

for the studied model and prove for it a stronger result. We prove that it is regular for all points out of the bifurcation diagram, i. e. for all non-critical values of the energy-momentum map.

Another condition of this type stated by V. Arnold and J. Moser (see [1, App. 8]) is that of the isoenergetical non-degeneracy which we explain further. Let us fix an energy level $H_0 = h_0$. If we get the Hamiltonian H_0 in action variables, then we can define the following map F_{h_0} from the $(n-1)$-dimensional variety $H_0^{-1}(h_0)$ into the projective space $\mathbf{P}^{n-1}$:

$$F_{h_0} : \ I \to (\omega_1(I) : \ldots : \omega_n(I)).$$

If the map F_{h_0} is a local diffeomorphism, we call this condition an isoenergetical non-degeneracy. Analytically, the isoenergetical non-degeneracy conditions are

$$\det \begin{pmatrix} \dfrac{\partial^2 H_0}{\partial I^2} & \dfrac{\partial H_0}{\partial I} \\ \dfrac{\partial H_0}{\partial I} & 0 \end{pmatrix} \neq 0. \tag{1.2}$$

Some years ago the potentials of the form of an S^1 symmetric double well were of interest to field theorists studying the Higgs field. In the present paper we study this condition for a model of a particle moving in a potential field shaped like the bottom of a bottle and determine thoroughly the set where it is violated for any energy level. It turns out to be either empty or consisting of two points. Of course, again the measure of the surviving tori depends on the measure of the set, where the above determinant is too close to zero.

Usually, it is difficult to check the conditions (1.1) and (1.2).

As far as I know, it has only been established for the spherical pendulum (see [4, 5]), Neumann's system, the geodesic flow on the ellipsoid (see [6]). The Kolmogorov condition for the Kirchhoff Top was proved in [9]. The condition of isoenergetical non-degeneracy for the problem of two centres of gravitation was checked in the paper [8]. We shall give the conditions (1.1) and (1.2) in terms of Abelian integrals and reduce the problem (as in [4, 5]) to analysis of these reminiscent and the study of limit cycles problems (see [7]).

2. THE ACTION VARIABLES

In this chapter we introduce some notations which we need in order to state the problem. We follow [2] and [4].

Let (M, ω) be a symplectic manifold of dimension $2n$, i.e. M is a smooth manifold and ω is a closed differential form of rank n. Let H be a smooth function on M. Denote by X_H the Hamiltonian vector field corresponding to the Hamiltonian H. Let also $f_1 \dots f_n$ be n functions in involution, i. e.

$$\{f_j,\ f_i\} = X_{f_j} f_i = 0, \qquad j, i = 1, \dots, n.$$

Define the level set

$$M_c = \{m\ :\ f_j(m) = c_j,\ j = 1, \dots, n\},$$

and suppose that the differentials are linearly independent on M_c. The following theorem gives complete description of the manifolds M_c together with the natural co-ordinates near them.

Theorem 2.1 (Liouville – Arnold). *Suppose $\tilde{M}_c$ is a compact component of M_c. Then:*

a) *$\tilde{M}_c$ is invariant under the flows generated by X_{f_j}, $j = 1, \dots, n$;*

b) *there are a neighbourhood U of $\tilde{M}_c$ and a diffeomorphism $J : f(U) \to V$, so that we have $I = J \circ f$, and the symplectic form ω in the co-ordinates (I, φ) takes a Darboux canonical form:*

$$\omega = \sum d\varphi \wedge dI. \tag{2.1}$$

(See [1] for the proof.) Recall that I, φ are called action-angle co-ordinates.

Following [2] and [4], one can construct the action co-ordinates. Let (p, q) be local Darboux co-ordinates such that the level surfaces $q_j = \text{const}$ meet transversally $\tilde{M}_c$. We suppose that the two-form ω is exact, $\omega = d\sigma$, where σ is an one-form.

Define a basis of cycles $\gamma_j(c)$, $j = 1, \ldots, n$, in the homology group $H(\tilde{M}_c, \mathbf{Z})$. Then the action variables are given by

$$I_k = \oint_{\gamma_k(c)} \sigma, \qquad k = 1, \ldots, n. \tag{2.2}$$

We define a model using a potential in the plane by

$$V(r) = r^4 - r^2, \tag{2.3}$$

where $r^2 = x^2 + y^2$ and x and y are the Cartesian co-ordinates in $\mathbf{R}^2$. The Hamiltonian of a particle moving in the plane under the influence of this potential is

$$H = \frac{1}{2}\left(p_x^2 + p_y^2\right) + \left(x^2 + y^2\right)^2 - \left(x^2 + y^2\right) \tag{2.4}$$

in the usual canonical co-ordinates (x, y, p_x, p_y). We change (2.4) into polar co-ordinates

$$x = r\cos\theta, \quad y = r\sin\theta.$$

Introducing the corresponding momenta $p_r = p_x$ and $p_\theta = p_y/r^2$, we obtain the Hamiltonian in the form

$$H = \frac{1}{2}\left(p_r^2 + \frac{1}{r^2}p_\theta^2\right) + r^4 - r^2. \tag{2.5}$$

Now $dp_\theta/dt = \{p_\theta, H\} = 0$, since θ is cyclic. Hence $G = p_\theta$ is the conserved angular momentum. This means that the Hamiltonian system is completely integrable, because we have the two conserved quantities G and H, whose Poisson brackets vanish.

We want to understand the geometry of the map J from $P = \mathbf{R}^4$ (the phase space) to $\mathbf{R}^2$, which is given by

$$J : P \to \mathbf{R}^2 \ : \ (x, y, p_x, p_y) \to (g, h),$$

where $H = h$.

The critical values of the map J are $(0, 0)$ and the curve is parameterized by

$$(g, h) = \left(\pm\sqrt{4r^6 - 2r^4}\ ,\ 3r^4 - 2r^2\right), \quad r \geq 2^{-1/2}$$

(see [2] for proofs). Denote by U_r the set of regular points of the map J (Fig. 1). For points $(g, h) \in U_r$ the level surface determined by the equations $H = h$, $G = g$ is a torus $T_{g,h}$. Choose a basis γ_1, γ_2 of the homology group $H_1(T_{g,h}, \mathbf{Z})$ with the following representations: for γ_1 take the curve on $T_{g,h}$, defined by fixing r and p_r and letting θ run through $[0, 2\pi]$; for γ_2 fix θ and p_r and let r and p_θ make one circle on the curve by the equation

$$h = \frac{1}{2}\left(p_r^2 + \frac{1}{r^2}p_\theta^2\right) + r^4 - r^2. \tag{2.6}$$

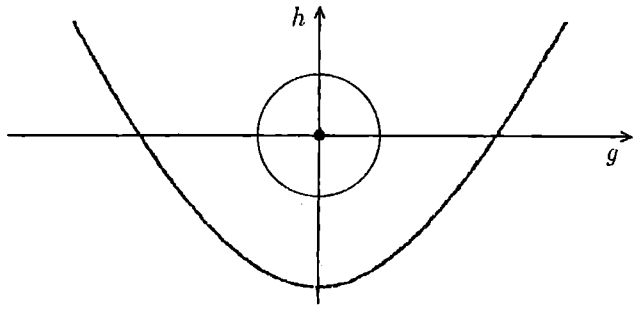


Fig. 1. Image of the map J

Now we can define the action co-ordinates I_1, I_2 by the formula (2.2), where

$$\sigma = p_\theta \wedge d\theta + p_r \wedge dr, \qquad \omega = d\sigma = dp_\theta \wedge d\theta + dp_r \wedge dr. \tag{2.7}$$

We have

$$I_1 = \oint_{\gamma_1} p_\theta \, d\theta = 2\pi g, \tag{2.8}$$

$$I_2 = \oint_{\gamma_2} p_r \, dr = 2\int_{r_1}^{r_2} \sqrt{2\left(h + r^2 - r^4 - \frac{g^2}{2r^2}\right)}\, dr, \tag{2.9}$$

where $r_1 < r_2$ are the roots of the equation $p_r = 0$ (see [2] and [4]). Put

$$z = r^2, \quad y = p_r r, \quad y^2 = 2(hz + z^2 - z^3) - g^2. \tag{2.10}$$

Denote the oval of the curve

$$\Gamma = \{(y, z) \ : \ y^2 = 2(hz + z^2 - z^3) - g^2\} \tag{2.11}$$

(which exists for all $(g, h) \in U_r$) by γ. Then we have

$$\psi(h, g) = I_2 = \int_\gamma \frac{y}{z}\, dz. \tag{2.12}$$

Let us show what is the meaning of r_1 and r_2. If the polynomial $P(z) = -2z^3 + 3z^2 + 2hz - g^2$ has three real roots $z_1 < z_2 < z_3$, then to r_1 corresponds z_2, and to r_2 corresponds z_3 (Fig. 2) in the proimage transformation (2.10).

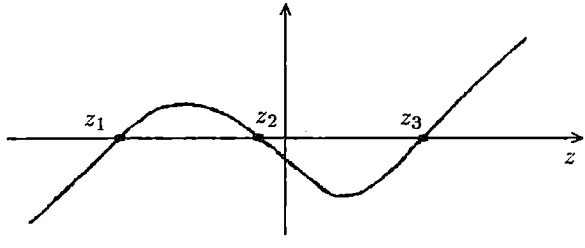


Fig. 2. Image of $P(z)$

Lemma 2.2. *The polynomial*

$$P(z) = -2z^3 + 3z^2 + 2hz - g^2$$

has three real different roots for all $(g, h) \in U_r$.

3. STATEMENT OF THE MAIN RESULT

Denote by $\tilde{H}(I_1, I_2)$ the Hamiltonian of our model in action co-ordinates. Our primary aim is to state the next theorem.

Theorem 3.1. *For* $(g, h) \in U_r$ *the determinant*

$$\det \begin{pmatrix} \dfrac{\partial^2 \tilde{H}}{\partial I_1^2} & \dfrac{\partial^2 \tilde{H}}{\partial I_1 \partial I_2} \\ \dfrac{\partial^2 \tilde{H}}{\partial I_2 \partial I_1} & \dfrac{\partial^2 \tilde{H}}{\partial I_2^2} \end{pmatrix} \tag{3.1}$$

does not vanish.

The condition (3.1) introduced by Kolmogorov [3] is crutial in KAM-theory [1, 3], dealing with the existence of invariant tori for perturbations of integrable systems. The procedure by which the invariant tori are constructed excludes the points, where the determinant (3.1) is violated, together with their neighbourhoods, whose measure is proportional to the perturbation (see [1]).

We shall give the condition (3.1) an explicite form in terms of Abelian integrals of the second kind. Using expression for I_1, I_2, we can determine $\tilde{G}$, $\tilde{H}$ implicitly from the equations

$$I_1 = 2\pi\tilde{G}, \quad I_2 = \psi(\tilde{G}, \tilde{H}). \tag{3.2}$$

Lemma 3.2. *The following formula holds true:*

$$(2\pi)^2 \left(\frac{\partial \psi}{\partial h}\right)^4 \det \begin{pmatrix} \dfrac{\partial^2 \tilde{H}}{\partial I_1^2} & \dfrac{\partial^2 \tilde{H}}{\partial I_1 \partial I_2} \\ \dfrac{\partial^2 \tilde{H}}{\partial I_2 \partial I_1} & \dfrac{\partial^2 \tilde{H}}{\partial I_2^2} \end{pmatrix} = \det \begin{pmatrix} \dfrac{\partial^2 \psi}{\partial^2 h} & \dfrac{\partial^2 \psi}{\partial h \partial g} \\ \dfrac{\partial^2 \psi}{\partial g \partial h} & \dfrac{\partial^2 \psi}{\partial g^2} \end{pmatrix}. \tag{3.3}$$

(For the proof see [4].)

Using [7], we have

$$\frac{\partial \psi}{\partial h} = \int_\gamma \frac{dz}{y} \neq 0 \tag{3.4}$$

in U_r.

Lemma 3.3. *For all* $(g,h) \in U_r$ *the determinant*

$$D = \det \begin{pmatrix} \dfrac{\partial^2 \psi}{\partial^2 h} & \dfrac{\partial^2 \psi}{\partial h \partial g} \\ \dfrac{\partial^2 \psi}{\partial g \partial h} & \dfrac{\partial^2 \psi}{\partial g^2} \end{pmatrix} \neq 0 \, .$$

This condition is equivalent to Theorem 1.

We formulate the condition of isoenergetical non-degeneracy in the next theorem.

Theorem 3.4. 1) *For* $h \in (-1/4, 0) \cup \left(\left(7\sqrt{249} - 1\right)/600, +\infty \right)$ *the map*

$$F_h : H^{-1}(h) \cap U_r \to \mathbf{P}^1, \quad F_h(I_1, I_2) = (H_{I_1} : H_{I_2})$$

is regular everywhere;

2) *For* $h \in \left(0, (7\sqrt{249} - 1)/600\right]$ *the map* F_h *has exactly two critical points.*

Next we would like to show that the entries of D can be represented as elliptic integrals. If we differentiate $\psi(h,g)$ twice formally, we get the following expressions:

$$\frac{\partial^2 \psi}{\partial h^2} = -\int\limits_\gamma \frac{z\,dz}{y^3}, \tag{3.5}$$

$$\frac{\partial^2 \psi}{\partial h \partial g} = g \int\limits_\gamma \frac{dz}{y^3}, \tag{3.6}$$

$$\frac{\partial \psi}{\partial g} = -g \int\limits_\gamma \frac{dz}{zy},$$

$$\frac{\partial^2 \psi}{\partial g^2} = -\int\limits_\gamma \frac{dz}{zy} - g \int\limits_\gamma \frac{g\,dz}{zy^2} = -\int\limits_\gamma \frac{(y^2+g^2)}{zy^3}\,dz = -2\int\limits_\gamma \frac{h+z+z^2}{y^3}\,dz\,. \tag{3.7}$$

The differential forms containing y^{-3} have poles along γ. There is a standard way to get rid of the poles on the integration path and we remind it below. Consider $\Gamma^{\mathbf{C}}_{g,h}$ as an elliptic curve in $\mathbf{C}$ defined by the equation for $\Gamma_{g,h}$. Topologically, it is a torus, whose one point is removed (see [4]). Now we deform the cycle γ on $\Gamma^{\mathbf{C}}_{g,h}$ into a new cycle γ' (Fig. 3) on which the function y has no zeroes. Of course, during the deformation the differential form $yz^{-1}\,dz$ must have no poles. Then by Cauchy's theorem the function $\psi(g,h)$ can be defined by the integral (2.12), taken on the path of integration γ' instead γ. With this definition of $\psi(g,h)$ the derivatives are well defined. We denote again γ' by γ.

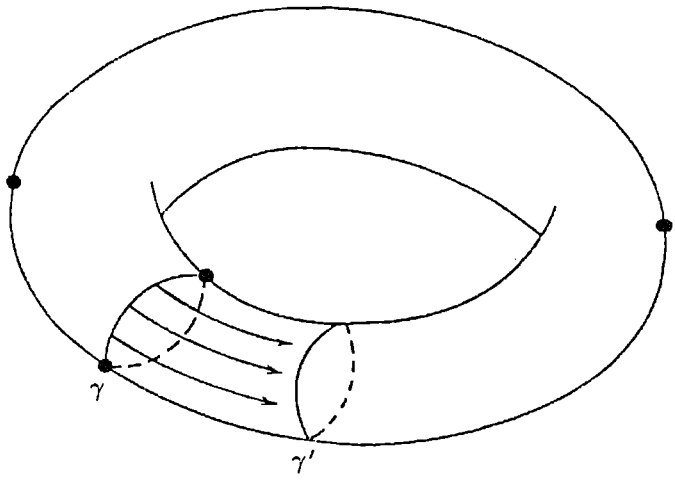


Fig. 3. The deformation of the cycle γ

Let

$$w_j(g,h) = \int_\gamma \frac{z^j}{y^3}\,dz, \qquad j = 0, 1. \tag{3.8}$$

The next lemma gives a representation of D as a quadratic form in w_0, w_1, which we shall need throughout this paper.

Lemma 3.5. *The determinant D has the representation*

$$D = \frac{2}{3} w_1(2hw_0 + w_1) - g^2 w_0^2. \tag{3.9}$$

Proof. We have

$$\frac{\partial^2 \psi}{\partial h^2} = -w_1$$

(see (3.5)),

$$\frac{\partial^2 \psi}{\partial h \partial g} = \frac{\partial^2 \psi}{\partial g \partial h} = g w_0$$

(see (3.6)),

$$\frac{\partial^2 \psi}{\partial g^2} = -2hw_0 - w_1 + 2\int_\gamma \frac{z^2}{y^3}\,dz$$

(see (3.7)). We need an expression for

$$2\int_\gamma \frac{z^2}{y^3}\,dz.$$

Let transform this integral in the following way: we have

$$2z^3 = 2(hz + z^2) - g^2 - y^2,$$

$$\begin{aligned}\int_\gamma \frac{2z^2}{y^3}\,dz &= \frac{1}{3}\int_\gamma \frac{2z^3}{y^3}\,dz = \frac{1}{3}\int_\gamma \frac{d(2(hz+z^2)-g^2-y^2))}{y^3} \\ &= \frac{1}{3}\int_\gamma \frac{2h}{y^3}\,dz + \frac{4}{3}\int_\gamma \frac{z}{y^3}\,dz - \frac{2}{3}\int_\gamma \frac{y}{y^3}\,dy = \frac{2h}{3}w_0 + \frac{4}{3}w_1,\end{aligned}$$

because

$$\int_\gamma \frac{dy}{y} = 0.$$

Then

$$\frac{\partial^2\psi}{\partial g^2} = -2hw_0 - 2w_1 + \frac{2h}{3}w_0 = \frac{4}{3}w_1 = -\frac{4h}{3}w_0 - \frac{2}{3}w_1,$$

this gives the representation (3.9).

We see that D does not depend on the sign of g. That is why it is enough to prove Lemma 3.3 only for $g \geq 0$.

4. PICARD–FUCHS EQUATIONS

Lemma 4.1. *Let $g = 0$. Then the functions w_0 and w_1 satisfy the following system of Picard–Fuchs equations:*

$$2h(4h+1)\frac{dw_0}{dh} = -2(7h+2)w_0 + 5w_1, \tag{4.1}$$

$$2(4h+1)\frac{dw_1}{dh} = w_0 - 10w_1. \tag{4.2}$$

Proof. Differentiating the expression (3.8) with respect to h, we obtain

$$\frac{dw_k}{dh} = -3\int_\gamma \frac{z^{k+1}}{y^5}\,dz, \quad k = 0, 1. \tag{4.3}$$

Put $g = 0$. Then we transform w_0 in the following way:

$$\begin{aligned}w_0 &= \int_\gamma \frac{dz}{y^3} = \int_\gamma \frac{y^2}{y^5}\,dz = 2\int_\gamma \frac{(hz + z^2 - z^3)}{y^5}\,dz \\ &= -\frac{2h}{3}\frac{dw_0}{dh} - \frac{2}{3}\frac{dw_1}{dh} - 2\int_\gamma \frac{z^3}{y^5}\,dz,\end{aligned}$$

$$
\begin{aligned}
2\int_\gamma \frac{z^3}{y^5}\,dz &= \frac{1}{3}\int_\gamma \frac{z}{y^5}\,d2z^3 = \frac{1}{3}\int_\gamma \frac{z}{y^5}\,d(2(hz+z^2)-y^2) \\
&= \frac{2h}{3}\int_\gamma \frac{z}{y^5}\,dz + \frac{4}{3}\int_\gamma \frac{z^2}{y^5}\,dz - \frac{2}{3}\int_\gamma \frac{z}{y^4}\,dy \\
&= -\frac{2h}{9}\frac{dw_0}{dh} - \frac{4}{9}\frac{dw_1}{dh} + \frac{2}{9}\int_\gamma z\,dy^{-3} \\
&= -\frac{2h}{9}\frac{dw_0}{dh} - \frac{4}{9}\frac{dw_1}{dh} - \frac{2}{9}\int_\gamma \frac{dz}{y^3} = -\frac{2h}{9}\frac{dw_0}{dh} - \frac{4}{9}\frac{dw_1}{dh} - \frac{2}{9}w_0.
\end{aligned}
$$

Then

$$
w_0 = -\frac{2h}{3}\frac{dw_0}{dh} - \frac{2}{3}\frac{dw_1}{dh} + \frac{2h}{9}\frac{dw_0}{dh} + \frac{4}{9}\frac{dw_1}{dh} + \frac{2}{9}w_0.
$$

This gives

$$
w_0 = -\frac{4h}{7}\frac{dw_0}{dh} - \frac{2}{7}\frac{dw_1}{dh}\,. \tag{4.4}
$$

In the same manner we transform w_1 and obtain

$$
w_1 = -\frac{2h}{35}\frac{dw_0}{dh} - \frac{28h+8}{35}\frac{dw_1}{dh}. \tag{4.5}
$$

Now solving (4.4) and (4.5), for $\dfrac{dw_0}{dh}$ and $\dfrac{dw_1}{dh}$ we get the system (4.1) and (4.2).

We also need the function

$$
\sigma(h) = \frac{w_1(h,0)}{w_0(h,0)}\,. \tag{4.6}
$$

Lemma 4.2. *The function $\sigma(h)$ satisfies the Riccati's equation*

$$
2h(4h+1)\frac{d\sigma}{dh} = -5\sigma^2 + 4(h+1)\sigma + h. \tag{4.7}
$$

Proof. Obviously,

$$
\frac{d\sigma}{dh} = \frac{1}{w_0^2}\left(w_0\frac{dw_1}{dh} - w_1\frac{dw_0}{dh}\right) = \frac{1}{2h(4h+1)}(-5\sigma^2 + 4(h+1)\sigma + h).
$$

When $g = 0$, the expression for D factors is

$$
D = \frac{2}{3}w_0^2\sigma\sigma_1, \tag{4.8}
$$

where $\sigma_1 = \sigma + 2h$.

From σ_1 we obtain the Riccati's equation

$$2(4h+1)\frac{d\sigma_1}{dh} = -5\sigma_1^2 + 4(6h+1)\sigma_1 + 8h^2 - 3h. \tag{4.9}$$

We need also some other functions both for the study of σ and σ_1, and for the case $g \neq 0$. In order to introduce them, we put the family of curves $\Gamma_{g,h}$ into the normal form

$$\Gamma_p = \{(u,v) \in \mathbf{C} : \ v^2 = 2(u^3 - 3u + p), \ p \in (-2,2)\}$$

by the transformation $z = -t + \frac{1}{3}$, $y = \alpha v$, $t = \beta u$, $\alpha = \beta^{3/2}$, where

$$\beta = \frac{1}{3}\sqrt{3h+1}\,, \tag{4.10}$$

$$p = \frac{1}{\beta^3}\left(\frac{h}{3} + \frac{2}{27} - \frac{g^2}{27}\right), \tag{4.11}$$

$p \in (-2,2)$ (see [7]). In these variables the integrals $w_0(g,h)$, $w_1(g,h)$ become

$$w_0 = -\frac{\beta}{\alpha^3}\int\limits_{\gamma(p)} \frac{du}{v^3}\,, \tag{4.12}$$

$$w_1 = -\frac{\beta}{\alpha^3}\int\limits_{\gamma(p)} \frac{-\beta u + (1/3)}{v^3}\,du. \tag{4.13}$$

We introduce the new functions

$$\theta_0(p) = \int\limits_{\gamma(p)} \frac{du}{v^3}\,, \quad \theta_1(p) = \int\limits_{\gamma(p)} \frac{u\,du}{v^3} \tag{4.14}$$

and their ratio

$$\varrho(p) = \frac{\theta_1(p)}{\theta_0(p)}\,. \tag{4.15}$$

In these notations we have

$$\sigma(h) = -\beta\varrho(p(0,h)) + \frac{1}{3}\,.$$

Lemma 4.3. 1) *The functions $\theta_0(p)$, $\theta_1(p)$ satisfy the Picard–Fuchs system*

$$6(4-p^2)\frac{d\theta_0}{dp} = 7p\theta_0 + 10\theta_1, \tag{4.16}$$

$$6(4-p^2)\frac{d\theta_1}{dp} = 14\theta_0 + 5p\theta_1. \tag{4.17}$$

2) *The function $\varrho(p)$ satisfies the Riccati's equation*

$$3(4-p^2)\frac{d\varrho}{dp} = 7 - p\varrho - 5\varrho^2. \tag{4.18}$$

The proof is the same as the one of Lemma 4.1 (see [4]).

5. ASYMPTOTIC BEHAVIOUR

Lemma 5.1. *The following formulas hold true:*

$$\lim_{p\to 2} \varrho(p) = 1, \tag{5.1}$$

$$\lim_{p\to -2} \varrho(p) = \frac{7}{5}, \tag{5.2}$$

$$\lim_{h\to -\frac{1}{4}} \sigma(h) = \frac{1}{10}, \tag{5.3}$$

$$\lim_{h\to 0} \sigma(h) = 0, \tag{5.4}$$

$$\lim_{h\to +\infty} \sigma(h) = -\infty, \tag{5.5}$$

$$\lim_{h\to +\infty} \frac{\sigma(h)}{h} = 0. \tag{5.6}$$

Proof. The proof of (5.1) and (5.2) is given in [4]. To prove (5.3)–(5.6), note that

$$\lim_{h\to -\frac{1}{4}} p(0,h) = -2, \qquad \lim_{h\to 0} p(0,h) = 2.$$

Then we obtain

$$\lim_{h\to -\frac{1}{4}} \sigma(h) = -\lim_{h\to -\frac{1}{4}} \beta \lim_{p\to -2} \varrho(p) + \frac{1}{3} = \frac{1}{10}.$$

Next we have

$$\lim_{h\to 0} \sigma(h) = -\frac{1}{3} + \frac{1}{3} = 0,$$

$$\lim_{h\to +\infty} \frac{\sigma(h)}{h} = -\lim_{h\to +\infty} \frac{\beta}{h} \lim_{p\to -2} \varrho(p) + \frac{1}{3} \lim_{t\to +\infty} = 0.\frac{7}{5} + 0 = 0.$$

And finally,

$$\lim_{h\to +\infty} \sigma(h) = -\lim_{h\to +\infty} \beta \lim_{p\to -2} \varrho(p) + \frac{1}{3} = -\infty.$$

Let us first consider the case $g = 0$.

Lemma 6.1. *The functions $\sigma(h)$, $\sigma_1(h)$ satisfy the following inequalities:*
1) *in the region* $-\frac{1}{4} < h < 0$, $\sigma(h) > 0$ and $\sigma_1(h) < 0$;
2) *in the region* $0 < h < +\infty$, $\sigma(h) < 0$ and $\sigma_1(h) > 0$.

Proof. First we prove that $\sigma(h)$ is positive in the interval $\left(-\frac{1}{4}, 0\right)$ and negative in $(0, +\infty)$. Let $h \in \left(-\frac{1}{4}, 0\right)$ and suppose that h_1 is the first zero of $\sigma(h)$ in this region. Then, using the Riccati's equation(4.7), we have

$$\sigma'(h_1) = \frac{1}{2(4h_1 + 1)} > 0.$$

The function $\sigma'(h)$ is continuous. That is why we obtain that a neighbourhood of point h_1 exists, where $\sigma'(h) > 0$. Then the function $\sigma(h)$ is strictly increasing in this neighbourhood. Using(5.3), we obtain that a point $h_0 < h_1$ exists, where $\sigma(h_0) = 0$: an obvious contradiction. In the same manner we obtain that $\sigma(h)$ can have no zero in the interval $(0, +\infty)$. Using Lemma 5.1, we obtain that $\sigma(h) > 0$ for $h \in \left(-\frac{1}{4}, 0\right)$ and $\sigma(h) < 0$ for $h \in (0, +\infty)$ (see Fig. 4). In the same way we obtain that the function $\sigma_1(h)$ is negative in the interval $\left(-\frac{1}{4}, 0\right)$.

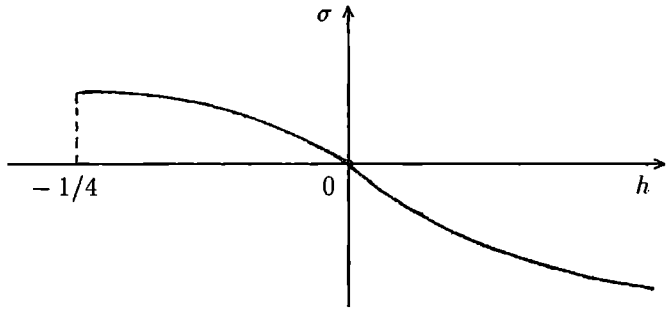


Fig. 4. Image of $\sigma(h)$

In order to proof that $\sigma_1(h) > 0$, we need the next proposition.

Lemma 6.2. *The function $\varrho(p)$ is decreasing on the interval $(-2, 2)$ and*

$$1 < \varrho(p) < \frac{7}{5}. \tag{6.1}$$

(For proof see [4].)

We have

$$\sigma_1(h) = -\beta\varrho(p(0,h)) + \frac{1}{3} + 2h > -\beta + \frac{1}{3} + 2h. \tag{6.2}$$

Using(4.10) and the substitution $\sqrt{3h+1} = t$, where $t \in (1, +\infty)$ for $h \in (0, +\infty)$, for the right hand side of (6.2) we obtain the new function

$$\eta(t) = 2t^2 - t - 1.$$

We shall prove that $\eta(t) > 0$ for $t \in (1, +\infty)$. Indeed,

$$\eta'(t) = 4t - 1\ ,$$

that is why the function $\eta(t)$ is strictly increasing on the interval $(1, +\infty)$. Now we have

$$\eta(t) > \eta(1) = 0.$$

We obtain that $\sigma_1(h) > 0$ for $h \in (0, +\infty)$. This completes the proof of Lemma 6.1 (see Fig. 5).

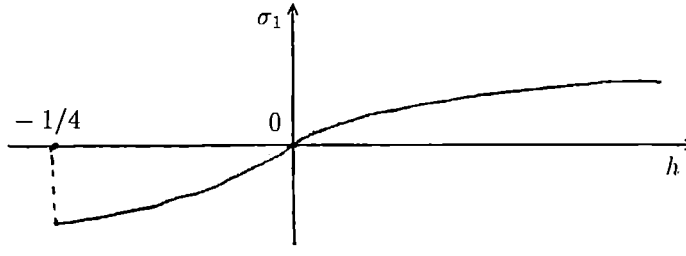


Fig. 5. Image of $\sigma_1(h)$

Corollary 6.3. *D is negative for $g = 0$.*

We turn to the general case $g > 0$.

Lemma 6.4. 1) *For $h \in \left(-\frac{1}{4}, 0\right) \cup (0, +\infty)$ and $g > 0$ we have the representation*

$$D = \frac{2}{3} w_0^2 \beta^2 . F(p, \beta),$$

where

$$F(p, \beta) = \varrho^2 - 6\beta\varrho + 3\beta p - 2. \tag{6.3}$$

2) *The functions $\beta(h)$ and $p(h, g)$ map the set*

$$U_r \bigcap \left\{ (g, h) :\ h \in \left(-\frac{1}{4}, 0\right) \cup (0, +\infty) \right\}$$

diffeomorphically on the set

$$V_r = \left\{(p,\beta):\ \beta \in \left(\frac{1}{6},\frac{1}{3}\right) \bigcup \left(\frac{1}{3},\ +\infty\right),\ p \in (-2,\,2)\right\}.$$

Proof. For D we have (using (3.9), (4.10) and (4.11))

$$\begin{aligned} D &= \frac{2}{3}w_1\,(2hw_0 + w_1) - g^2 w_0^2 = \frac{1}{3}w_0^2\left(2\left(-\beta\varrho + \frac{1}{3}\right)\left(2h - \beta\varrho + \frac{1}{3}\right) - 3g^2\right) \\ &= \frac{1}{3}w_0^2\left(2\beta^2\varrho^2 - 12\beta^3\varrho + 2\beta^2 - \frac{2}{9} - 3g^2\right) \\ &= \frac{1}{3}w_0^2\left(2\beta^2\varrho^2 - 12\beta^3\varrho - \frac{2}{3}(3h+1) + 2\beta^2 + 6\beta^2 p\right) \\ &= \frac{1}{3}w_0^2\left(2\beta^2\varrho^2 - 12\beta^3\varrho - 4\beta^2 + 6\beta^3\right) = \frac{2}{3}w_0^2\beta^2\left(\varrho^2 - 6\beta\varrho + 3\beta p - 2\right). \end{aligned}$$

Lemma 6.5. *For all $(p,b) \in V_r$ the function F is negative.*

Proof. We have

$$\frac{\partial F}{\partial \beta} = -6\varrho + 3p, \qquad \frac{\partial^2 F}{\partial\beta\partial p} = -6\varrho' + 3 > 0,$$

because $\varrho' < 0$ (see Lemma 6.2).

That is why we obtain that the function $\dfrac{\partial F}{\partial\beta}$ is a strictly increasing function of $p \in (-2,\,2)$. Now we have

$$\frac{\partial F}{\partial\beta}(p,\beta) < \frac{\partial F}{\partial\beta}(2,\beta) = -6\varrho(2) + 3.2 = 0,$$

then $F(p,\beta)$ is a strictly decreasing function of β $(\beta \geq \dfrac{1}{6})$. We obtain

$$F(p,\beta) < F\left(p,\frac{1}{6}\right) = \varrho^2 - \varrho + \frac{p}{2} - 2,$$

but $-1 < \dfrac{p}{2} < 1$ and $-\dfrac{7}{5} < -\varrho < -1$. So now we obtain

$$-\varrho + \frac{p}{2} < 0, \qquad \varrho^2 < \frac{49}{25}, \qquad \varrho^2 - 2 < -\frac{1}{25},$$

hence $F(p,\beta) < 0$. This completes the proof of Lemma 6.5 and together with that the proof of Theorem 3.1.

Our aim is the proof of Theorem 3.4. Here we find an expression for the function F_h in the terms of elliptic integrals. We have $F_h = F_h(g)$, $h = \text{const}$.

Lemma 7.1. *Let $(g,h) \in U_r$. Then F_h has the representation*

$$F_h(g) = -\frac{1}{2\pi}\frac{\partial\psi}{\partial g}(g,h) = g\int_\gamma \frac{dz}{zy}. \tag{7.1}$$

The proof is straightforward.

Lemma 7.1 shows that we have to determine the zeroes of the function

$$\frac{\partial^2\psi}{\partial g^2}(g,h) = -2\frac{\partial}{\partial g}F_h(g)$$

for a fixed h. We shall study the curve of zeroes of the function $\dfrac{\partial^2\psi}{\partial g^2}(g,h)$ for $(g,h) \in U_r$. The statement of the theorem easily follows from the properties of this curve. Because of the symmetry of the set U_r with respect to the line $g = 0$, we concentrate our attention on the set $U^+ = U_r \cup \{g \geq 0\}$.

Lemma 7.2. *For $g = 0$ and $(0,h) \in U_r$ the function $\dfrac{\partial^2\psi}{\partial g^2}$ does not vanish.*

The proof is a simple application of Lemma 6.1 and (3.4).

Now let $g \neq 0$. It is clear that we study only the case $g > 0$. We have

$$\frac{\partial^2\psi}{\partial g^2} = \frac{2}{3}w_0\left(2h + \frac{w_1}{w_0}\right) = \frac{2}{3}w_0\beta\left(\varrho - 6\beta + \frac{1}{3\beta}\right).$$

We know that $\beta \neq 0$, that is why we obtain the equation

$$\varrho - 6\beta + \frac{1}{3\beta} = 0, \quad \beta \in \left(\frac{1}{6},\frac{1}{3}\right) \cup \left(\frac{1}{3},+\infty\right), \tag{7.2}$$

$$\varrho = 6\beta - \frac{1}{3\beta}, \quad 1 < \varrho < \frac{7}{5}.$$

Then we get

$$\beta \in \left(\frac{1}{3}, \frac{7+\sqrt{249}}{60}\right].$$

Proof of Theorem 3.4. Let $\beta \in \left(\dfrac{1}{3}, \dfrac{\sqrt{249}+7}{60}\right]$. Then the equation(7.2) has exactly one solution $p(\beta) \in [-2,2]$, as Lemma 6.2 implies. This defines a function $\beta \to p(\beta)$, $\beta \in \left(\dfrac{1}{6},\dfrac{1}{3}\right) \cup \left(\dfrac{1}{3},+\infty\right)$, which is strictly increasing. Our aim is to

prove that the curve in U^+, defined as the zero-locus of the function $\frac{\partial^2\psi}{\partial g^2}$, has exactly one point of intersection with the line $h = h_0$ for $h_0 \in \left(0, \frac{7+\sqrt{249}}{600}\right]$ (the image of the interval $\beta \in \left(\frac{1}{3}, \frac{7+\sqrt{249}}{60}\right]$ by (4.10)). Suppose there are two points g_1 and g_2 for which

$$\frac{\partial^2\psi}{\partial g^2}(g_j, h_0) = 0, \quad j = 1, 2.$$

Then the images of these points (g_j, h_0) by the transformation (4.10), (4.11), which we denote by (p_j, b_0), $j = 1, 2$, satisfy the equation (7.2) for $\beta_0 \in \left(\frac{1}{3}, \frac{\sqrt{249}+7}{60}\right]$. Because of $\varrho(p)$ being strictly increasing, we obtain $p_1 = p_2$. But $g > 0$ and using (4.10) we have $g_1 = g_2$. This finishes the proof of Theorem 3.4.

ACKNOWLEDGEMENTS. The author thanks Emil Horozov for the helpful discussions.

The research is partially supported by the Ministry of Science and Technology under Contract N 523/95.

REFERENCES

1. Arnold, V. I. Mathematical Methods of Classical Mechanics. Berlin – Heidelberg – New York, 1978.
2. Bates, L. Monodromy in the champagne bottle. *ZAMP*, 1, 1991.
3. Kolmogorov, A. N. On the preservation of the conditionally periodic motions under small perturbations of the Hamiltonian functions. *Dokl. Akad. Nauk. SSSR*, **98**, 1954.
4. Horozov, E. Perturbations of the spherical pendulum and Abelian integrals. *J. reine angew. Math.*, **408**, 1990.
5. Horozov, E. On the isoenergetical non-degeneracy of the spherical pendulum. *Phisics Letters A*, **173**, 1993.
6. Knörer, H. Singular fibers of the momentum mapping for integrable Hamiltonian systems. *J. reine angew. Math.*, **355**, 1985.
7. Chow, S. N., J. A. Sanders. On the number of critical points of the period. *J. Diff. Equ.*, **64**, 1986.
8. Dragnev, D. On the isoenergetical non-degeneracy of the Problem of two centers of gravitation. *Phisics Letters A*, **215**, 1996.
9. Christov, O. On the Kolmogorov's condition for the Kirchhoff Top. *J. Math. Phys.*, **37**, 1994.

Received on July 25, 1996

Differential Equations
Faculty of Mathematics and Informatics
Sofia University
5 James Bourchier
1164 Sofia, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 1 — Математика и механика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 1 — Mathématiques et Mecanique
Tome 89, 1995

ON THE "TRIANGULAR" INEQUALITY IN THE THEORY OF TWO-PHASE RANDOM MEDIA

KONSTANTIN Z. MARKOV

A necessary condition on the two-point correlation function of binary random media, noticed by Matheron [1] and called by him "triangular" inequality, is studied in this note. An appropriate result, due to Achiezer and Glazman [2], is first recalled. Simple consequences of this inequality are given, as well as a necessary condition for its validity in a statistically isotropic medium. It is shown that it represents a requirement, independent of that of the familiar positive definiteness, that should be additionally imposed on the two-point correlation function of any realistic binary medium.

Key words: random materials, two-phase media, correlation functions.
1991/95 Mathematics Subject Classification: 60G60, 73B35.

Consider a random and statistically uniform medium that occupies d-dimensional space $\mathbb{R}^d$. The medium is "binary", i.e., it consists of two phases labelled 1 and 2. Phase 1 (which needs not to be connected) occupies Ω_1 and phase 2 occupies its complement Ω_2. The characteristic function of Ω_1 is f_1. Thus,

$$f_1(x) = \begin{cases} 1, & \text{if} \quad x \in \Omega_1, \\ 0, & \text{otherwise.} \end{cases}$$

As it is well-known, the statistical properties of the medium follow from the set of multipoint probabilities or moments of f_1:

$$\eta_1 = \langle f_1(0) \rangle , \quad \langle f_1(0) f_1(z_1) \rangle , \ldots , \tag{1}$$

where each $z_k \in \mathbb{R}^d$, see for instance [3]. The angled brackets signify ensemble

averaging. Such multipoint probabilities are symmetric in their arguments. One point could be taken at the origin, because of the assumed statistical uniformity.

It is, in fact, convenient to work with η_1 and the multipoint moments

$$M_p(z_1, z_2, \ldots, z_{p-1}) = \langle f_1'(0) f_1'(z_1) f_1'(z_2) \ldots f_1'(z_{p-1}) \rangle, \quad p = 2, 3, \ldots, \tag{2}$$

where

$$f_1'(z) = f_1(z) - \eta_1 \tag{3}$$

is the fluctuating part of the field $f_1(z)$.

Of course, not any infinite hierarchy of functions M_p can represent moments derived from a random medium and, moreover, from a two-phase one. The reason, well recognized and very clearly explained by Frisch [4], is that the function M_p should satisfy, in particular, certain compatibility conditions. The real problem in this connection arises when modelling a random constitution of practical interest. In such cases the first few moments (as a rule the two-point and, more rarely, the three-point ones) are prescribed using certain, very often heuristic and not very rigorous arguments. Though the form of the prescribed moments can, in principle, be checked experimentally, the question remains as to whether these moments can be inserted into the infinite hierarchy of multipoint moments (2), i.e. whether they pertain to a *real* random medium. The problem is even tougher when the two-phase media are dealt with, having in mind that the latter very often appear in application. Frisch [4], for example, presented examples of two-point probability densities that look plausible but *cannot* belong to any real random medium. Another more recent example is connected with the often used "well-stirred" approximation for random dispersion of spheres, for which, as far as the two-point moment is only concerned, overlapping is forbidden and the sphere location is not statistically interconnected otherwise. This approximation turns out to be realistic only at sphere fraction $\eta_1 \leq 1/8$ in 3D, as shown in [5, 6].

For any statistically homogeneous medium one restriction that is generally known is that its two-point correlation function should be positive definite, so that its Fourier transform must be positive. The converse is also true, namely, for any positive-definite function there exists a random medium for which this function represents its two-point correlation (the Bochner or Bochner-Khinchine theorem, see, e.g., [3]). Further restrictions are known if the medium is also statistically isotropic [3]. For two-phase media, as introduced above, it *ought* to be possible to find more restrictions but none are known; a conjecture on how to recognize realistic two-point correlation functions for such media was recently made by Matheron [1]. As a matter of fact, a method for deriving relations of such a type has been proposed in the recent work [6] on the basis of a certain variational reasoning.

Here we shall study in more detail a requirement, specific for the correlation of a two-phase medium. This is an inequality first noticed, to the best of the author's knowledge, by Matheron [1] and called by him "triangular" due to obvious geometrical reasons. It appears that this inequality closely resembles a certain property of the positive definite functions, first pointed out by Achiezer and Glazman [2] almost forty years ago. That is why we shall first recall the appropriate result of Achiezer and Glazman.

Following these authors, introduce the class $\mathcal{G}$ of real and even functions $g(x)$, $x \in \mathbb{R}^d$, for which the kernel

$$\Gamma(x,y) = g(x) + g(y) - g(x-y) \tag{4}$$

is positive definite, i.e.

$$\sum_{i,j=1}^{k} \Big[g(x_i) + g(x_j) - g(x_i - x_j) \Big] a_i a_j \geq 0, \quad \forall x_i \in \mathbb{R}^d,\ a_i \in R. \tag{5}$$

Proposition 1. *Let $\gamma_2(x)$ be a real positive definite and even function on $\mathbb{R}^d$. Then $1 - \gamma_2(x) \in \mathcal{G}$ (and thus $\lambda(1-\gamma_2(x)) \in \mathcal{G}$ as well, $\forall \lambda \geq 0$).*

Proof. Due to the definition (5), $1 - \gamma_2(x) \in \mathcal{G}$ if the kernel

$$T(x,y) = 1 + \gamma_2(x-y) - \gamma_2(x) - \gamma_2(y) \tag{6}$$

is positive definite. To prove this, consider the identity

$$\sum_{i,j=1}^{2k} \gamma_2(y_i - y_j) b_i b_j = \sum_{i,j=1}^{2k} \gamma_2(y_{2i} - y_{2j-1}) b_{2i} b_{2j-1}$$

$$+ \sum_{i,j=1}^{2k} \gamma_2(y_{2i} - y_{2j}) b_{2i} b_{2j} + \sum_{i,j=1}^{2k} \gamma_2(y_{2i-1} - y_{2j-1}) b_{2i-1} b_{2j-1}$$

$$+ \sum_{i,j=1}^{2k} \gamma_2(y_{2i-1} - y_{2j}) b_{2i-1} b_{2j}.$$

Choose now $y_{2i} = 0$, $y_{2i-1} = x_i$, $b_{2i} = -a_i$, $b_{2i-1} = a_i$, $i = 1, \ldots, k$. Then

$$0 \leq \sum_{i,j=1}^{2k} \gamma_2(y_i - y_j) b_i b_j = \sum_{i,j=1}^{k} \Big[1 + \gamma_2(x_i - x_j) - \gamma_2(x_i) - \gamma_2(x_j) \Big] a_i a_j.$$

Hence the kernel $T(x,y)$, see (6), is indeed positive definite, which proves the proposition.

Remark 1. The Proposition 1 and its simple proof, given here for the sake of completeness, belong to Achiezer and Glazman [2], see also [7, p. 265].

Let the medium be two-phase and let

$$\gamma(x', x'') = \gamma(x' - x'') = \frac{1}{2} \Big\langle |f_1(x') - f_1(x'')|^2 \Big\rangle \tag{7}$$

denote the so-called variogramme of the field $f_1(x)$. Using the definition of the two-point correlation, it is easily seen that

$$\gamma(x) = \eta_1 \eta_2 (1 - \gamma_2(x)), \tag{8}$$

where

$$\gamma_2(x) = \frac{M_2(x)}{M_2(0)} = \frac{\langle f_1'(0) f_1'(x) \rangle}{\langle f_1'^2(0) \rangle}, \quad M_2(0) = \langle f_1'^2(0) \rangle = \eta_1 \eta_2,$$

so that $\gamma_2(x)$ is the most often used two-point correlation for which $\gamma_2(0) = 1$.

According to Proposition 1, $\gamma \in \mathcal{G}$, since $\gamma_2(x)$ is positive definite. Hence the field $\Gamma(x, y)$, generated by $\gamma(x)$, see (4), is positive definite. The following proposition shows, however, that for a two-phase medium an additional fact holds.

Proposition 2. *The variogramme of any two-phase random medium generates a field* $\Gamma(x, y)$ *which is not only positive definite, but which is nonnegative itself. In other words, the so-called* triangular inequality *of Matheron* [1] *holds:*

$$\gamma(x - y) \leq \gamma(x) + \gamma(y), \quad \forall x, y \in \mathbb{R}^d. \tag{9}$$

Proof. Obviously,

$$\gamma(x, y) = \frac{1}{2}\Big\langle |f_1(x) - f_1(y)|^2 \Big\rangle = \frac{1}{2}\Big\langle |f_1(x) - f_1(0) + f_1(0) + f_1(y)|^2 \Big\rangle$$

$$= \frac{1}{2}\Big\langle |f_1(x) - f_1(0)|^2 \Big\rangle + \frac{1}{2}\Big\langle |f_1(0) - f_1(y)|^2 \Big\rangle - \alpha(x, y)$$

$$= \gamma(x) + \gamma(y) - \alpha(x, y),$$

where

$$\alpha(x, y) = \Big\langle (f_1(0) - f_1(x))(f_1(0) - f_1(y)) \Big\rangle.$$

To prove (9) it suffices to show that $\alpha(x, y) \geq 0$. But, if the origin 0 lies in the constituent '2', then $f_1(0) = 0$ and $\alpha(x, y) = \langle f_1(x) f_1(y) \rangle \geq 0$. Similarly, if 0 lies in the constituent '1', then $f_1(0) = 1$ and again $\alpha(x, y) = \langle (1 - f_1(x))(1 - f_1(y)) \rangle \geq 0$.

Combining (8) and (9) yields

$$\gamma_2(x) + \gamma_2(y) - \gamma_2(x - y) \leq 1 \tag{10}$$

or

$$1 + \gamma_2(r' + \theta r'') \geq \gamma_2(r') + \gamma_2(r''), \quad \forall \theta \in [-1, 1], \tag{11}$$

having chosen $|x| = r'$, $r'' = |y|$. This inequality should thus be satisfied by the two-point correlation of *any* realistic statistically homogeneous and two-phase random medium.

Corollary 1. *Let the medium be statistically isotropic as well, so that* $\gamma_2(x) = \gamma_2(r)$, $r = |x|$. *Then*

$$\gamma_2'(0) \leq_{\cdot} \pm \gamma_2'(r), \quad \forall r \in (0, \infty). \tag{12}$$

Indeed, choose the vectors x, y colinear, once with the same directions and then with the opposite directions; $|y| = \Delta r$, $|x| = r$, $r > \Delta r > 0$. Then

$$\gamma_2(\Delta r) + \gamma_2(r) \leq 1 + \gamma_2(r \pm \Delta r)$$

which, at $\Delta r \ll 1$, implies (12).

Since $\gamma_2(0) = 1$ and $\gamma_2(r) \leq 1$, we have obviously $\gamma_2'(0) \leq 0$. The inequality (12) is then equivalent to

$$|\gamma_2'(r)| \leq |\gamma_2'(0)|, \quad \forall r \in (0, \infty), \tag{13}$$

which means, in particular, that the steepest decrease of the two-point correlation function $\gamma_2(r)$ of an isotropic two-phase medium is at the origin $r = 0$.

Corollary 2. *A positive definite function $\gamma_2(r)$ may serve as a two-point correlation of a two-phase statistically homogeneous and isotropic medium, only if $\gamma_2'(0) < 0$.*

Indeed, (13) immediately shows that $\gamma_2'(0) = 0$ yields $\gamma_2'(r) = 0$, $\forall r \in (0, \infty)$, i.e. $\gamma_2(r) \equiv 1$, which is impossible.

The inequality $\gamma_2'(0) < 0$ for a two-phase medium follows also from the fact that $-\gamma_2'(0)$ is proportional to S/V, where S is the specific surface (i.e. phase boundary) within the small volume V, see [8] and especially [9, p. 177] for details and a proof. More precisely, $S/V = -4\eta_1(1 - \eta_1)\gamma_2'(0)$, which obviously implies $\gamma_2'(0) < 0$ for such media.

Remark 2. As it is well-known, not every real and positive function is positive definite and vice versa. Hence the triangular inequality represents a necessary condition that should be imposed on the two-point correlations of random media in *addition* to their positive definiteness, if the modelled medium is two-phase. To illustrate this consider as an example first the function

$$\gamma_2(r) = \frac{1}{(1 + (r/\beta)^2)^2}.$$

It is positive definite (since its Fourier transform is positive) and hence it represents, according to the Bochner-Khinchine theorem, a two-point correlation of a certain statistically homogeneous and isotropic random medium. On the other hand, $\gamma_2'(0) = 0$, so that the triangular inequality fails for this medium and the latter therefore *cannot* be two-phase.

Conversely, consider again the above mentioned "well-stirred" dispersion of spheres. Its two-point correlation satisfies the triangular inequality for all values of the sphere fraction $\eta_1 \in (0, 1)$ (since the field $f_1(x)$ is binary). On the other hand, the Fourier transform of this correlation is positive definite only at $\eta_1 \leq 1/8$, as it can be directly shown. Hence the positive definiteness and triangular inequality are indeed two mutually independent necessary conditions that should be satisfied by the two-point correlation of binary random media.

In general, it seems hard to give a complete description of the functions that satisfy the triangular inequality (10). (The variogrammes under study cannot obviously be homogeneous of degree 1, i.e. $\gamma(\lambda x) \neq \lambda\gamma(x)$, and thus they are not semi-norms on $\mathbb{R}^d$.) A simple and rich class of such function can be easily described though. To this end note that (11) implies $\gamma_2''(0) \geq 0$ for such a function

and thus $\gamma_2(r)$ is convex and monotonically increasing in a certain vicinity of the origin. If the latter properties hold for all $r \in [0, \infty)$, it suffices to claim that the respective function is an admissible two-point correlation. More precisely, recall that in 1D a bounded even function which is convex on the right half-axis is positive definite [10, p. 187]. A radially symmetric function $\gamma_2(r)$ in 3D with these properties is not obliged to be of that kind.[1] However, for such functions the following result holds:

Proposition 3. *If $\gamma_2(r)$ is monotonically decreasing, nonnegative and convex, then it satisfies the triangular inequality* (10).

Proof. Since $\gamma_2(r)$ is monotonically decreasing, in order to prove (10) it suffices to show that

$$1 + \gamma_2(r' + r'') \geq \gamma_2(r') + \gamma_2(r''),$$

having taken the vectors x, y colinear, with the same direction; $r' = |x|$, $r'' = |y|$. Let $r' > r''$ for definiteness. Then

$$1 - \gamma_2(r'') = \gamma_2(0) - \gamma_2(r'') = -\gamma_2'(\xi')r'', \quad \xi' \in (0, r''),$$

$$\gamma_2(r') - \gamma_2(r' + r'') = -\gamma_2'(\xi'')r'', \quad \xi'' \in (r', r' + r'').$$

The convexity of $\gamma_2(r)$ means that $\gamma_2''(r) \geq 0$, so that $\gamma_2'(\xi'') \geq \gamma_2'(\xi')$, because $\xi'' > \xi'$. Hence

$$1 - \gamma_2(r'') \geq \gamma_2(r') - \gamma_2(r' + r''),$$

which proves the proposition.

A simple example of an admissible and physically reasonable two-point correlation is

$$\gamma_2(r) = e^{-\mu r}, \tag{14}$$

proposed by Debye *et al.* [8]. This is the so-called exponential correlation, discussed, for instance, in the book of Stoyan *et al.* [9] (where a planar random set with this correlation is explicitly constructed in Sec. 10.5.1). Being convex, positive and monotonically decreasing, the function (14) satisfies the triangular inequality (10), as it follows from Proposition 3. Its Fourier transform is positive. Hence this function may represent a two-point correlation for a two-phase statistically homogeneous and isotropic medium in $\mathbb{R}^d$ for any d. A more general class of similarly admissible correlations is obviously given by

$$\gamma_2(r) = \int_0^\infty e^{-rt}\, d\sigma(t); \tag{15}$$

[1] The function in 3D

$$\gamma_2(r) = \begin{cases} 1 - r/a, & \text{if } r \leq a, \\ 0, & \text{if } r > a, \end{cases}$$

$r = |x|$, is bounded and convex, but its Fourier transform $\widehat{\gamma}_2(k)$ is proportional to $2(1 - \cos ak) - ak \sin ak$ and hence it is not positive everywhere.

here $\sigma(t)$ is an arbitrary bounded and non-increasing function on $(0,\infty)$ such that $\int_0^\infty d\sigma(t) = 1$. (If $\sigma(t) = H(t-\mu)$, Debye's function (14) is recovered from (16), $H(t)$ being the Heaviside function.) In other words, the class (16) gathers the Laplace transforms of all nonnegative functions on $(0,\infty)$ (more precisely, of all bounded measures there).

Note finally that the class (15) coincides with the class of the so-called completely monotonic functions, according to the well-known Bernstein theorem, see, for instance, [11] or [7]. It is curious, however, that such completely monotonic functions (15) may represent correlations only for dispersions of overlapping or touching particles. The reason is that non-overlapping always implies the condition $\gamma_2''(0) = 0$, as it follows from the results of Kirste and Porod [12], see also [13] and [5]. This condition, however, never holds for the functions (15).

Another example of an admissible two-point correlation is the function

$$\gamma_2(r) = \begin{cases} 1 - \dfrac{3r}{4a} + \dfrac{r^3}{16a^3}, & \text{if } r \le 2a, \\ 0, & \text{if } r > 2a, \end{cases} \tag{16}$$

since it is obviously positive definite, nonnegative and convex. Hence it satisfies the triangular inequality (10) as well. Note that (16) is the two-point correlation of the so-called Miller's cell material [14] in the simplest case when the cells are spherical, see also [15].

ACKNOWLEDGEMENTS. The support of this work by the Bulgarian Ministry of Education, Science and Technology under Grant No MM416-94 is gratefully acknowledged.

REFERENCES

1. Matheron, G. Une conjecture sur covariance d'un ensemble aleatoire. *Cahiers de Géostatistique*, Fasc. **3**, 1993, 107–113.
2. Achiezer, N. I., I. M. Glazman. On certain classes of continuous functions generating Hermitian-positive kernels. *Soobshcenija Harkovskogo Mat. Obshtestva*, Harkov, vol. **25**, 1957 (in Russian).
3. Vanmarcke, E. Random fields: Analysis and synthesis. MIT Press, Cambridge, Massachusetts and London, England, 1983.
4. Frisch, H. L. Statistics of random media. *Trans. Soc. Rheology*, **9**, 1965, 293–312.
5. Markov, K. Z. On a statistical parameter in the theory of random dispersions of spheres. In: Continuum Models of Discrete Systems, Proc. 8^{th} Int. Symposium (Varna, 1995), K. Z. Markov, ed., World Sci., 1996, 241–249.
6. Markov, K. Z., D. R. S. Talbot, J. R. Willis. On stationary diffusion in heterogeneous media. *IMA J. Applied Mathematics*, **56**, 1996, 133–144.
7. Achiezer, N. I. The classical problem of moments. Gos. Izd. Fiz.-Mat. Lit., Moscow, 1961 (in Russian).
8. Debye, P., H. R. Anderson, Jr. H. Brumberger. Scattering by an inhomogeneous solid. II. The correlation function and its application. *J. Appl. Phys.*, **28**, 1957, 679–683.

9. Stoyan, D., W. S. Kendall, J. Mecke. Stochastic geometry and its applications. Akademic-Verlag Berlin, GDR/John Wiley & Sons Ltd., 1987.
10. Donogue, W. F., Jr. Distributions and Fourier transforms. Academic Press, New York and London, 1969.
11. Widder, D. V. The Laplace transform. Princeton University Press, Princeton, N.J., 1946.
12. Kirste, R., G. Porod. Röntgenkleinwinkelstreuung an kolloiden Systemen, Asymptotisches Verhalten der Streukturven. *Kolloid-Z.*, **184**, 1962, 1–7.
13. Frisch, H. L., F. H. Stillinger. Contribution to the statistical geometric basis of radiation scattering. *J. Chem. Phys.*, **38**, 1963, 2200–2207.
14. Miller, M. N. Bounds for effective electrical, thermal and magnetic properties of heterogeneous materials. *J. Math. Phys.*, **10**, 1969, 1988–2004.
15. Hori, M. Statistical theory of effective electrical, thermal, and magnetic properties of random heterogeneous materials. I. Perturbation expansions for the effective permittivity of cell materials. *J. Math. Phys.*, **14**, 1973, 514–523.

Received on September 17, 1996

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5 blvd J. Bourchier
BG-1164 Sofia, Bulgaria
e-mail: kmarkov@fmi.uni-sofia.bg

ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Книга 2 — ПРИЛОЖНА МАТЕМАТИКА И ИНФОРМАТИКА

Том 89
1995

ANNUAIRE

DE

L'UNIVERSITE DE SOFIA "ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE

Livre 2 — MATHEMATIQUES APPLIQUEE ET INFORMATIQUE

Tome 89
1995

СОФИЯ • 1998 • SOFIA
УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
PRESSES UNIVERSITAIRES "ST. KLIMENT OHRIDSKI"

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 89, 1995

FOURIER-GALERKIN ALGORITHM FOR 2D LOCALIZED SOLUTIONS

(Dedicated to the memory of my young colleague K. L. Bekyarov)

CHRISTO I. CHRISTOV

The paper presents the numerical implementation in 2D of a Fourier-Galerkin expansion with complete orthonormal basis system of localized functions. The bilinear Laplace equation is considered as a featuring example. Coordinate splitting is used to reduce the cost of inversion of the linear matrices for the coefficients. The axisymmetric soliton is calculated as a 2D problem and compared to a numerical solution, found by means of a difference scheme.

Keywords: Fourier-Galerkin method, localized solutions, bilinear Laplace equation.
1991 Mathematics Subject Classification: 65N30.

1. INTRODUCTION

Calculating the shapes of localized waves, e.g. solitons, is of importance for the modern theory of non-linear waves. The difficulties are connected with the unboundness of the integration domain. For example, in numerical treatment, when using finite-difference or finite-element schemes, one has to consider large enough domains in order to reduce the influence of the truncation (the so-called "actual infinity"). In 1D the problems of domain size and mesh resolution can still be tackled, although sometimes up to 20000 grid points (see, e.g. [12]) have to be used. Clearly, in 2D, when the mesh size is at least the square of the 1D mesh-size, it is a very hard problem.

One of the ways to circumvent the said difficulty is to employ a complete orthonormal (CON) system of functions on the infinite interval and to devise an

algorithm for implementation of one of the spectral techniques: Galerkin's, pseudospectral, *tau*-method (see [5, 3]). The successful application of the Galerkin method requires, however, that the product of two members of the system can be conveniently represented by means of the functions of the system. CON system with the required properties was introduced first in [6] and applied for finding a localized solution to the Burgers equation. Later on, the numerical Fourier-Galerkin technique was extended to Korteweg-de Vries (KdV) and Kuramoto-Sivashinsky (KS) equations [11] and the fifth order KdV [1]. Boyd [2, 4] showed that the new CON system can be obtained by an algebraic mapping of the Tchebishev polynomials on an infinite interval, see also [3]. In this way he derived a variety of properties of the expansion.

Employing a spectral expansion with a specialized CON basis system drastically reduces the required computational resources. They can be even further reduced if the resulting algebraic system is treated in the appropriate manner by means of a splitting method. The aim of the present paper is the creation of an algorithm for implementing the Fourier-Galerkin technique in 2D.

2. POSING THE PROBLEM

Consider the following generic equation (the non-linear Klein-Gordon equation)

$$\frac{\partial^2 u}{\partial t^2} = -u + 3u^2 + \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{2.1}$$

which, as is well-known, possesses localized solutions that propagate stationary. In the co-ordinate system connected with the center of the localized structure (the so-called "moving frame") one can introduce new independent variables $\xi x - c_1 t$, $\eta = y - c_2 t$, where c_1, c_2 are the components of the phase speed of the center of the localized structure. Then for the stationary localized solution one arrives at the equation

$$-u + 3u^2 + \left(\beta_1 \frac{\partial^2 u}{\partial \xi^2} + \beta_2 \frac{\partial^2 u}{\partial \eta^2}\right) = 0, \tag{2.2}$$

where $\beta_i = 1 - c_i^2$. Here we consider only "subsonic" solitons for which $\beta_i > 0$. The boundary conditions stem from the vanishing of the solution at infinity:

$$u \to 0 \quad \text{for} \quad \xi, \eta \to \pm\infty. \tag{2.3}$$

Clearly, the problem (2.2), (2.3) is a bifurcation one, since the trivial solution $u \equiv 0$ always persists. A similar problem was treated in [14] for the classical spectral method with harmonic functions in application to the sixth order Boussinesq equation. To avoid the trivial solution, one can impose a condition at the origin of the co-ordinate system, say,

$$u(0,0) = \text{const}. \tag{2.4}$$

Strictly speaking, (2.4) will overpose the problem unless some additional degree of freedom is introduced, say, through an additional coefficient of the non-linear term

$$-u + 3\alpha u^2 + \left(\beta_1 \frac{\partial^2 u}{\partial \xi^2} + \beta_2 \frac{\partial^2 u}{\partial \eta^2}\right) = 0, \tag{2.5}$$

which is to be calculated so as to fit the imposed boundary conditions at the origin of the co-ordinate system. The definitive relation for the new unknown is the equation taken in the origin:

$$\alpha = \frac{1}{3u^2(0,0)} \left[u(0,0) - \left(\beta_1 \frac{\partial^2 u}{\partial \xi^2} + \beta_2 \frac{\partial^2 u}{\partial \eta^2}\right)\Big|_{x=0,\, y=0} \right]. \tag{2.6}$$

The last relation does not overpose the problem, since the equation in the origin is not used in the scheme for u, but rather it is replaced by the prescribed boundary condition (2.4). Thus we arrive to a boundary value problem (b.v.p.) which does not possess a trivial solution. In addition, for the unknowns (u, α) explicit relations are available. Then the construction of an iterative procedure is straightforward. In some cases, however, the convergence is achieved only when a relaxation for α is performed.

Note that the above procedure is valid only when the expected solution has non-zero amplitude in the origin of the co-ordinate system. When this is not the case (say, for solutions that are odd functions), one can impose a similar condition on one of the partial derivatives of u in the origin. In order not to overload the presentation, we skip the details of such a case and consider here only the case of even functions.

3. FOURIER-GALERKIN EXPANSION

3.1. THE BASIS SYSTEM OF FUNCTION IN $L^2[-\infty, \infty]$

The first CON system in $L^2(-\infty, \infty)$ suited for non-linear problems was proposed in [6]. The different formulas were compiled and verified in [7]. Here we cite the necessary formulas in order to make the paper self-content.

The products of members of series are expanded in series of the system

$$C_n C_k = \frac{1}{2\sqrt{2\pi}} [\ C_{n+k+1} - C_{n+k} - C_{n-k} + C_{n-k-1}] = \sum_{n=0}^{\infty} \beta_{nk,l} C_l, \tag{3.1}$$

$$S_n S_k = \frac{1}{2\sqrt{2\pi}} [\ C_{n+k+1} - C_{n+k} + C_{n-k} - C_{n-k-1}] = \sum_{n=0}^{\infty} \alpha_{nk,l} C_l, \tag{3.2}$$

$$S_n C_k = \frac{1}{2\sqrt{2\pi}} [-S_{n+k+1} + S_{n+k} + S_{n-k} - S_{n-k-1}] = \sum_{n=0}^{\infty} \gamma_{nk,l} S_l. \tag{3.3}$$

The first derivatives of the functions of the system are expressed as

$$\frac{dS_n}{dx} = \frac{1}{2}\left[nC_{n-1} + (2n+1)C_n + (n+1)C_{n+1}\right],$$

$$\frac{dC_n}{dx} = -\frac{1}{2}\left[nS_{n-1} + (2n+1)S_n + (n+1)S_{n+1}\right].$$

Respectively, for the second derivatives one has

$$\frac{d^2C_n}{dx^2} = -\frac{1}{4}\left\{n(n-1)C_{n-2} - 4n^2C_{n-1} + \left[n^2 + (n+1)^2 + (2n+1)^2\right]C_n\right.$$

$$\left. -4(n+1)^2C_{n+1} + (n+1)(n+2)C_{n+2}\right\}, \tag{3.4}$$

$$\frac{d^2S_n}{dx^2} = -\frac{1}{4}\left\{n(n-1)S_{n-2} - 4n^2S_{n-1} + \left[n^2 + (n+1)^2 + (2n+1)^2\right]S_n\right.$$

$$\left. -4(n+1)^2S_{n+1} + (n+1)(n+2)S_{n+2}\right\}. \tag{3.5}$$

3.2. THE GALERKIN EXPANSION

The simplest and oldest spectral technique is the Galerkin one in which the sets of test and trial functions coincide. The main purpose of the present work is to provide an efficient iterative algorithm for treating the linear part of the system. For this reason we select a system with a quadratic non-linearity, for which the Galerkin method is the most efficient. When a more complicated non-linearity is present, then one of the pseudo-spectral techniques should be used. In addition, our equation admits even solution. That is why, for the sake of simplicity, we consider the following series:

$$u = \sum_{n=0}^{n=N} a_{mn}C_m(x)C_n(y). \tag{3.6}$$

3.3. THE CONDITIONS FOR COUPLING THE SYSTEM

Introducing the expressions for the derivatives in the differential equation, one gets a five-diagonal system for each subsystem of coefficients C_n, S_n. The system has to be truncated at $n = 0$ (no terms of negative order show up, since they are expressed by the functions of positive order) and for certain sufficiently large $n = N$. Then the problem of coupling conditions arises. Here we resort to even functions only and the formulas are similar for the odd functions. The condition for coupling the system for $n = 0$ and $n = 1$ comes from the very formulae of the second derivatives (3.4)

$$\frac{d^2C_0}{dx^2} = -\frac{1}{2}C_0 + C_1 - \frac{1}{2}C_2, \tag{3.7}$$

$$\frac{d^2C_1}{dx^2} = C_0 - \frac{7}{2}C_1 + 4C_2 - \frac{3}{2}C_1. \tag{3.8}$$

In the framework of the Galerkin method, the truncation of the system at $n = N$ requires to assume that $C_n \equiv 0$ and $C_{n+1} \equiv 0$. Then, for the last two members of the series one gets the following expressions for their second derivatives:

$$\frac{d^2C_{n-1}}{dx^2} = -\frac{(n-2)(n-1)}{4}C_{n-3} + (n-1)^2C_{n-2} - \frac{3n^2-3n+1}{2}C_{n-1} + n^2C_n, \tag{3.9}$$

$$\frac{d^2C_n}{dx^2} = -\frac{n(n-1)}{4}C_{n-2} + n^2C_{n-1} - \frac{3n^2+3n+1}{2}C_n. \tag{3.10}$$

Thus the second x-derivative in the governing equation (2.2) is approximated by 5-point finite difference in the system. Denote by Λ_{xx} and Λ_{yy} the respective five-diagonal matrices which are obtained after half of the identity operator, $\frac{1}{2}$, is subtracted from each of the second-derivative operators. Then the original equation is approximated by the algebraic system

$$3\alpha \sum_{n_1=0}^{n_1=N} \sum_{m_1=0}^{m_2=M} \sum_{n_2=0}^{n_2=N} \sum_{m_2=0}^{m_2=M} \beta^x_{n_1m_1,n}\beta^y_{n_2m_2,m}a_{m_1n_1}a_{m_2n_2} + (\Lambda_{xx} + \Lambda_{yy})a_{mn} = 0, \tag{3.11}$$

where β^x and β^y are the matrices of coefficients from (3.1) for x and y, respectively. The system (3.11) is taken for all $n \neq 0$ and $m \neq 0$. In the origin the boundary condition

$$a_{00} = 1$$

is imposed. Respectively, the system (3.11), taken at $n = 0, m = 0$, gives the definitive relation for α, namely

$$\hat{\alpha} = \frac{-\beta_1\left(-\frac{1}{2}a_{00} + a_{10} - \frac{1}{2}a_{20}\right) - \beta_2\left(-\frac{1}{2}a_{00} + a_{01} - \frac{1}{2}a_{02}\right)}{3\sum_{n_1=0}^{n_1=N} \sum_{m_1=0}^{m_1=M} \sum_{n_2=0}^{n_2=N} \sum_{m_2=0}^{m_2=M} \beta^x_{n_1m_1,0}\beta^y_{n_2m_2,0}a_{m_1n_1}a_{m_2n_2}}, \tag{3.12}$$

where the unknowns a_{mn} are from the new iteration (fictitious-time stage) $k+1$. The relaxation for α is performed as follows:

$$\alpha^{k+1} \stackrel{\text{def}}{=} \alpha^k(1-\omega) + \hat{\alpha}\omega.$$

4. THE SPLITTING SCHEME

In previous works on 1D problems ([11, 1]) we used the Brent's routine for solving the non-linear system for the coefficients. Despite of the rather simple expressions for the products of members of system into series in the system (see

(3.1) — (3.3)), using a pseudo-Newton algorithm like the Brent's one becomes too expensive in 2D because of the large size of the Jacobian. This justifies the search for alternative algorithms. Here we use a simple iteration for the non-linear term. The appropriate series representation of the products of the terms in the system is rather "sparse," so a lot of iterations can be easily performed. It is desirable, however, to have the linear part approximated implicitly. We split it to reduce the calculations. Thus we use the following scheme corresponding to the so-called (see [15]) scheme of stabilizing correction:

$$\frac{\tilde{a}_{ij} - a^n_{ij}}{\tau} = \Lambda_{xx}\tilde{a}_{ij} + \Lambda_{yy}a^k_{ij} + F[a^k_{ij}] \tag{4.1}$$

$$\frac{a^{k+1}_{ij} - \tilde{a}_{ij}}{\tau} = \Lambda_{yy}[a^{k+1}_{ij} - a^k_{ij}]. \tag{4.2}$$

Here τ is the time increment with respect to the fictitious time and it plays the role of an iteration parameter. Respectively, $F[a^n]$ is the expression for the non-linear term when evaluated with the values for a_{ij} from the "old" iteration

$$F[a^k] = \sum_{n=0}^{n=N}\sum_{m=0}^{m=M} \beta^x_{n_1 m_1, n}\beta^y_{n_2 m_2, m}a^k_{mn}\,.$$

After excluding the half-time-step variable $\tilde{a}$, one gets

$$\left(E + \tau^2\Lambda_{xx}\lambda_{yy}\right)\frac{a^{k+1} - a^k}{\tau} = (\lambda_{xx} + \Lambda_{yy})a^{k+1} + F[a^k] \tag{4.3}$$

which converges to (3.11) in the limit $k \to \infty$, when $a^{k+1} \to a^k$. The important feature of the system (4.1), (4.2) is that it requires inversion of five-diagonal matrices for which special very fast elimination algorithms are available. We make use here of the algorithm from [9].

The iterations are terminated when the following criterion is satisfied

$$\left|a^{k+1} - a^k\right| < 10^{-10}.$$

5. RESULTS AND DISCUSSION

5.1. THE AXISYMMETRIC LOCALIZED SOLUTION

2D calculations of solitons are rarely found. That is why there are no available cases for comparison. However, for $\beta_1 = \beta_2$ one can compare a cross-section of the solution obtained by the 2D algorithm of the present work to 1D solution of the equation when the axial symmetry is acknowledged. Hence we consider the equation

$$u - 3u^2 - \frac{\beta_1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) = 0, \tag{5.1}$$

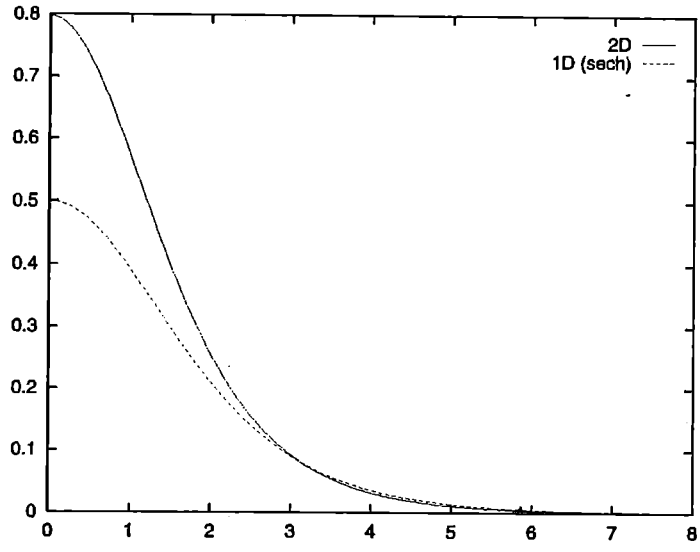


Fig. 1. The axisymmetric soliton for $\beta_1 = \beta_2 = 1$ as obtained with $N + 1 = 20$ functions in the spectral expansion

for which a localized solution is sought in $-\infty, \infty$. To this end we employ the so-called Method of Variational Imbedding (MVI), proposed in [8] for the homoclinic solution of the Lorenz system. To an equation of the type of (5.1), but with a cubic non-linearity, the MVI was applied in [10]. The algorithmic problems of application of MVI are elucidated in detail in [13] in application to the solitary-wave solution of the Kuramoto-Sivashinsky equation. For this reason we present here only the result for the axisymmetric soliton. Fig. 1 shows the shape of this solution alongside with the well-known *sech*-solution of the 1D case. It is seen that the axisymmetric soliton is taller (maximum height equal to 0.79735, while in 1D the maximum equals exactly 0.5) and of slightly smaller support. The solution presented in the figure is taken as a reference when assessing the approximation of the spectral scheme in the next subsection.

5.2. VERIFICATION OF ALGORITHM

The practical convergence of the method can be assessed if a cross section of the 2D solution is taken as function of the radial co-ordinate. Fig. 2 shows the result for different number of terms in the spectral series. Being reminded that the maximum of solution is approximately 0.8, one sees that even 8 functions are able to provide approximation closer to the solution than 0.3%, and 20 terms in the series give approximation better than 0.006%. It is to be mentioned here that no special care for optimization of the method has been taken in the present work. As shown in [11, 1], one can further improve the approximation with fewer number of terms by means of scaling the independent variable(s) in order to bring it closer to the characteristic measures (length of support) of the basis functions C_n, S_n.

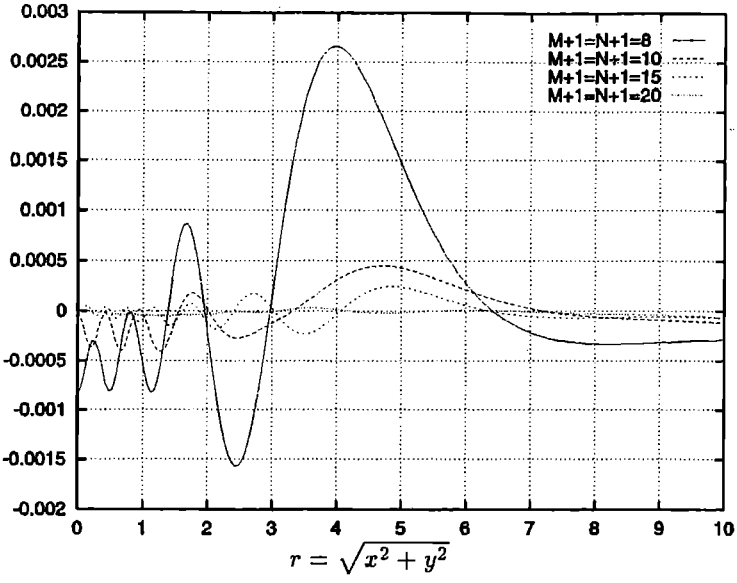


Fig. 2. The difference between the spectral solution with different number of functions and the finite difference solution with 401 points in the interval [0, 9.9875]

In two dimensions the shape of soliton is presented in Fig. 3 as obtained by the 2D algorithm developed here. Note that a cross-section of this solution is compared in Fig. 2 to the solution with radial symmetry from Fig. 1.

5.3. THE NON-AXISYMMETRIC SOLUTION

As mentioned in the precedence, the convergence of the spectral series can be improved ([11]) if one succeeds to select the optimal scaling for the independent variable. This is especially important when in two dimensions the coefficients before the different highest-order derivatives differ significantly. In our case these are the coefficients β_1 and β_2. The optimization needs a special attention together with an extensive set of numerical experiments and goes beyond the framework of the present paper. Here we have only demonstrated the effectiveness of the splitting scheme for solving the algebraic system for the coefficients. For this reason we do not scale the independent variables even for the case shown in Fig. 4, where there is a considerable difference between the two coefficients $\beta_2 = 1 = 10\beta_1$, $\beta_1 = 0.1$.

In this case a solution obtained by an independent numerical technique is not available and the convergence test is performed by the standard increase of the number of terms in the expansion and by assessing the contribution of the last term. Once again, employing 15 terms gives accuracy of 0.1% and 20 terms bring the difference down to 0.01%. This means that even for one order of magnitude difference between the coefficients of the second derivatives, 20 terms in the expansion is fully enough for securing a very good accuracy. When the ratio between the

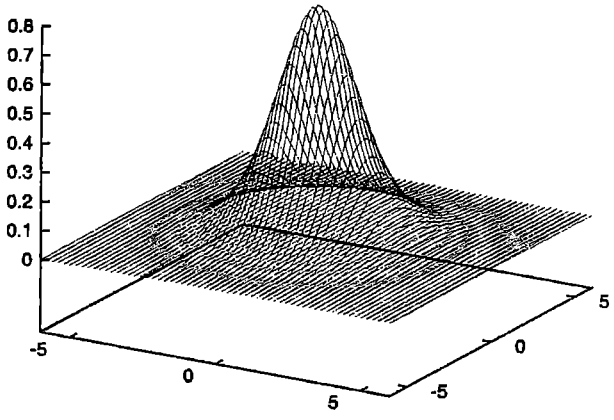


Fig. 3. The axisymmetric soliton for $\beta_1 = \beta_2 = 1$ as obtained with $M + 1 = N + 1 = 20$ functions in the spectral expansion

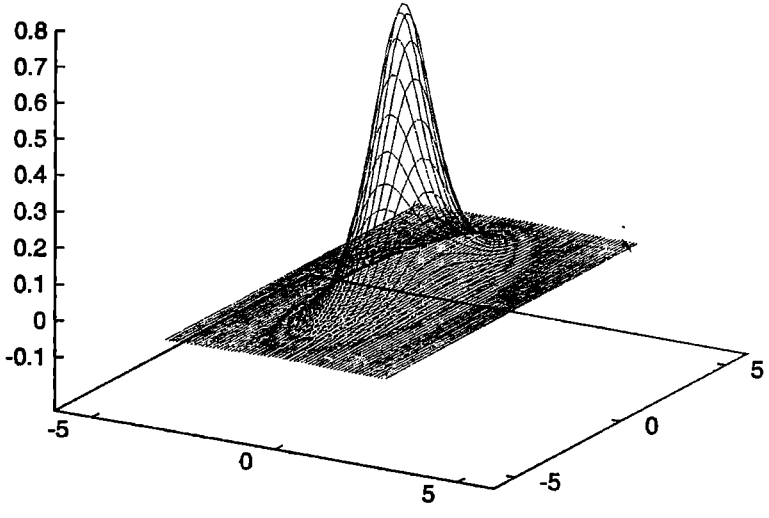


Fig. 4. The soliton for $\beta_1 = 0.1$, $\beta_2 = 1$ as obtained with $M + 1 = N + 1 = 20$ functions in the spectral expansion

coefficients β_i is still larger, one can attempt optimization of the algorithm through different scaling of the independent variables (see [11] for the details in 1D).

6. CONCLUSION

In the present paper a Fourier-Galerkin algorithm for numerical treatment of the bifurcation problem for localized solutions of 2D non-linear PDE is developed.

To avoid the always present trivial solution, an additional boundary condition is imposed in the origin of the co-ordinate system and a coefficient is added before the non-linear term. The equation itself taken in the origin serves as an explicit definitive relation for the new coefficient. The iterative procedure involves artificial time and co-ordinate splitting of the linear operator corresponding to the partial derivatives. The convergence is secured through selecting the values of the artificial-time increment and the relaxation parameter for the sought coefficient of the non-linear term. In 2D the splitting-type procedure has a significant advantage over the direct Newton-type quasi-linearization algorithms for solving the algebraic system for the coefficients of the Galerkin expansion.

Results are obtained for a generic equation of Klein-Gordon's type with a quadratic non-linearity. The 1D an axisymmetric soliton of the equation in the moving frame is obtained by means of two different techniques and the comparisons give the quantitative assessment of the truncation errors of the spectral expansion.

ACKNOWLEDGEMENTS. This work was partially supported by the National Science Foundation of Bulgaria under Grant NZ-611/96.

REFERENCES

1. Bekyarov, K. L., C. I. Christov. Fourier–Galerkin numerical technique for solitary waves of the fifth-order KdV equation. *Chaos, Solitons & Fractals*, **1**, 1992, 423–430.
2. Boyd, J. P. Spectral methods using rational basis functions on an infinite interval. *J. Comp. Phys.*, **69**, 1987, 112–142.
3. Boyd, J. P. Spectral Methods. Springer-Verlag, New York, 1989.
4. Boyd, J. P. The orthogonal rational functions of Higgins and Christov and algebraically mapped Chebishev polynomials. *J. Approx. Theory*, **85**, 1990.
5. Canuto, C., M. Y. Hussaini, A. Quarteroni, T. A. Zang. Spectral Methods in Fluid Dynamics. Springer, 1987.
6. Christov, C. I. A complete orthonormal sequence of functions in $l^2(-\infty, \infty)$ space. *SIAM J. Appl. Math.*, **42**, 1982, 1337–1344.
7. Christov, C. I. A method for treating the stochastic bifurcation of plane Poiseuille flow, nonlinear stochastic systems. *Ann. Univ. Sof., Fac. Math. Mech.*, **76**, Livre 2, Mécanique, 1982, 87–113.
8. Christov, C. I. A method for identification of homoclinic trajectories. In: *Proc. 14-th Spring Conf. Sunny Beach*, Sofia, Bulgaria, 1985, 571–577.
9. Christov, C. I. Gaussian elimination with pivoting for multidiagonal systems. *Internal Report*, **4**, University of Reading, 1994.
10. Christov, C. I. On the mechanics of localized structures in continuous media. In: *Fluid Physics*, eds. M. G. Velarde and C. I. Christov, World Scientific, 1995, 33–60.
11. Christov, C. I., K. L. Bekyarov. A Fourier–series method for solving soliton problems. *SIAM J. Sci. Stat. Comp.*, **11**, 1990, 631–647.

12. Christov, C. I., G. A. Maugin. An implicit difference scheme for the long-time evolution of localized solutions of a generalized Boussinesq system. *J. Comp. Phys.*, **116**, 1995, 39–51.
13. Christov, C. I., M. G. Velarde. On localized solutions of an equation governing Bénard-Marangoni convection. *Appl. Math. Modelling*, **17**, 1993, 311–320.
14. Steyt, Y., C. I. Christov, M. G. Velarde. Solitary-wave solutions of a generalized wave equation with higher-order dispersion. In: *Continuum Models and Discrete Systems*, ed. K. Z. Markov, World Scientific, 1996, 471–479.
15. Yanenko, N. N. Method of Fractional Steps. Gordon and Breach, 1971.

Received on September 23, 1996

National Institute of Meteorology and Hydrology
Bulgarian Academy of Sciences
BG-1184 Sofia, Bulgaria
e-mail: christo.christov@meteo.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 89, 1995

MOUVEMENT D'UNE SPHÈRE HOMOGÈNE DANS UN CYLINDRE HORIZONTAL AVEC UN MOMENT RÉSISTANT DE FROTTEMENT

SONIA DENEVA

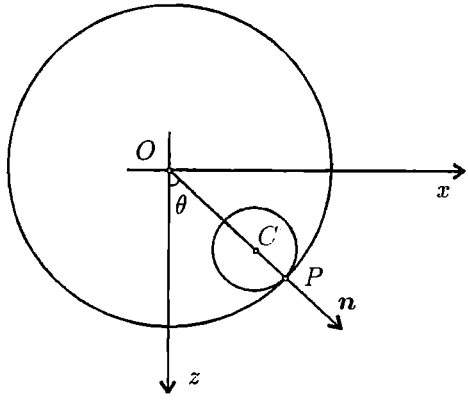
In this paper some aspects of the classical problem concerning rolling sphere on a homogeneous horizontal cylinder are considered.

Keywords: motion of a rigid sphere, friction.
1991/95 Mathematics Subject Classification: 70E15.

Soit donné un cylindere droit circulaire de rayon R posé en une position horizontale immobile. Avec le cylindre est lié un système de coordonnées $Oxyz$; pour Oy nous choisissons l'axe du cylindre et Oz a une direction verticale en bas. La sphère homogène de centre C, de masse m et de rayon r se roule sur la part inférieur du cylindre où elle a un point de contacte P. Nous supposons que le mouvement devient avec frottement entre deux corps mais si le coefficient du frottement est grand il n'est pas possible un mouvement avec glissement. Voila pourquoi nous supposons que le mouvement de la sphère est un roulement propre sans glissement mais nous prenons en considération qu'il y a un moment résistant de frottement contre roulement d'après Painlevé. Nous supposons encore que le plan équatorial de la sphère reste toujours sur le plan vertical Oxz du cylindre. La position de la sphère sur le cylindre est donnée au dessin 1. D'isi nous avons

$$\delta = R - r = OC. \tag{1}$$

Designons par θ l'angle $(\widehat{Oz, OP})$ qui détermine la position du point P. Le vecteur



Dess. 1

unique $\boldsymbol{n}$ qui est normal à deux surfaces s'exprime par la formule

$$\boldsymbol{n} = \sin\theta\,\boldsymbol{i} + \cos\theta\,\boldsymbol{k}. \tag{2}$$

Selon (1) et (2) nous obtenons

$$\begin{aligned} \overrightarrow{OC} &= \delta\sin\theta\,\boldsymbol{i} + \delta\cos\theta\,\boldsymbol{k}, \\ \boldsymbol{v}_C &= \delta\dot{\theta}(\cos\theta\,\boldsymbol{i} - \sin\theta\,\boldsymbol{k}), \end{aligned} \tag{3}$$

où $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$ sont orts du système $Oxyz$.

Puisque le plan équatorial de la sphère reste verticale, la vitesse angulaire ω de la sphère a la forme

$$\boldsymbol{\omega} = \omega\boldsymbol{j}, \quad \omega > 0. \tag{4}$$

Le mouvement est un roulement sans glissement et voila pourquoi nous avons la relation pour le point de contacte

$$\boldsymbol{v}_P = \boldsymbol{v}_C + \boldsymbol{\omega} \times \overrightarrow{CP} = 0. \tag{5}$$

Ayant vu (2), (3) et (4), nous obtenons de (5)

$$\omega = -\frac{\delta}{r}\dot{\theta} \tag{6}$$

(la grandeur $\dot{\theta} < 0$ parce que θ est une fonction de croissante du temps).

Nous montrons que tous les grandeurs de la cinématique et de la dynamique de la sphère peuvent s'exprimer comme des fonctions de θ.

Le théorème de la resultante cinétique et la théorème du moment cinétique appliqué au point C se traduisent par les équations

$$\frac{d}{dt}(m\boldsymbol{v}_C) = m\boldsymbol{g} + \boldsymbol{R}, \tag{7}$$

$$\frac{2}{5}mr^2\frac{d\boldsymbol{\omega}}{dt} = \overrightarrow{CP} \times \boldsymbol{R} + \boldsymbol{\Gamma}. \tag{8}$$

Ici $\boldsymbol{R}$ est la force de la résistance appliquée au point P; $\boldsymbol{\Gamma}$ est le moment résistant du frottement qui d'après Painlevé se donne par la formule

$$\boldsymbol{\Gamma} = -f'' R_{\boldsymbol{n}} \boldsymbol{j} = -f'' R_{\boldsymbol{n}} \frac{\boldsymbol{\omega}}{\omega}. \tag{9}$$

Le signe moins montre que le vecteur $\boldsymbol{\Gamma}$ a une direction inverse au vecteur $\boldsymbol{\omega}$, c'est-à-dire le vecteur $\boldsymbol{\Gamma}$ se résiste au roulement de la sphère déterminé par le vecteur $\boldsymbol{\omega}$. $R_{\boldsymbol{n}}$ est la grandeur de la projection de la force $\boldsymbol{R}$ sur le vecteur $\boldsymbol{n}$. Le coefficient f'' du frottement est ordinairement une petite grandeur. Puisque tous les vecteurs dans (6) se trouvent dans le plan Oxz, la force $\boldsymbol{R}$ a la forme

$$\boldsymbol{R} = R_x \boldsymbol{i} + R_z \boldsymbol{k}. \tag{10}$$

De l'équation (7) selon (10) on obtient

$$\begin{aligned} R_x &= m\delta \left(\ddot{\theta}^2 \cos\theta - \theta \sin\theta \right), \\ R_z &= -mg - m\delta \left(\ddot{\theta} \sin\theta + \dot{\theta}^2 \cos\theta \right). \end{aligned} \tag{11}$$

De l'équation (11) selon (2) on obtient

$$R_{\boldsymbol{n}} = mg \cos\theta + m\delta\dot{\theta}^2. \tag{12}$$

De l'équations (8) et (9) on obtient

$$\frac{2}{5} m r^2 \frac{d\omega}{dt} = -r \sin\theta R_z + r \cos\theta R_x - f'' R_{\boldsymbol{n}}. \tag{13}$$

Remplaçons dans (13) les relations (6), (11) et (12) et après quelques calculations on obtient

$$\theta = -\frac{5}{7}\frac{g}{\delta} \sin\theta - \frac{f''}{r}\dot{\theta}^2 + \frac{5}{7}\frac{f''}{\delta}\frac{g}{r} \cos\theta. \tag{14}$$

L'équation (14) nous pouvons ecrire à la forme

$$\frac{du}{d\theta} - \frac{10}{7}\frac{f''}{r} u = -\frac{10}{7}\frac{g}{\delta} \sin\theta + \frac{10}{7}\frac{f''}{\delta}\frac{g}{r} \cos\theta; \quad u = \dot{\theta}^2. \tag{15}$$

Ayant vu que le coefficient f'' est une grandeur petite, nous obtenons de (15)

$$u = \dot{\theta}^2 = C \left(1 + \frac{10}{7}\frac{f''}{r}\theta \right) + \frac{10}{7}\frac{g}{\delta} \cos\theta + \frac{170}{49}\frac{f''}{\delta}\frac{g}{r} \sin\theta, \tag{16}$$

où C est une constante qui dépend des conditions initiables. Puisque $\dot{\theta} < 0$ on obtient de (16)

$$\dot{\theta} = -\sqrt{C \left(1 + \frac{10}{7}\frac{f''}{r}\theta \right) + \frac{10}{7}\frac{g}{\delta} \cos\theta + \frac{170}{49}\frac{f''}{\delta}\frac{g}{r} \sin\theta}. \tag{17}$$

En fin nous obtenons de (6) et (17)

$$\omega(\theta) = \frac{\delta}{r}\sqrt{C \left(1 + \frac{10}{7}\frac{f''}{r}\theta \right) + \frac{10}{7}\frac{g}{\delta} \cos\theta + \frac{170}{49}\frac{f''}{\delta}\frac{g}{r} \sin\theta}.$$

Les grandeurs R_n, R_x, R_z s'expriment aussi par l'angle θ des formules (10) et (11).

BIBLIOGRAPHIE

1. Painlevé, P. Leçons sur le frottement. Moscou, 1954 (en Russe).
2. Cabannes, H. Cours de mecanique generale. Paris, 1962.
3. Deneva, S. Mouvement avec frottement d'une sphère homogène dans un cylindre horizontal. *Annuaire de l'Univ. de Sofia, Fac. de Math. et Informatique*, t. 87, 1993.

Received on June 15, 1996

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 89, 1995

COOPERATION OF CLIENT ROUTINES IN CLIENT-SERVER NETWORK ARCHITECTURE WITHOUT USING OF SPECIAL MONITOR ROUTINE ON SERVER

PETER DIMOV

A method for communication between client routines without monitoring by a special routine, working on a server, is proposed. The base of the method is a message transferring engine, but not in the classical form. The method is oriented to data sharing between computers in a client-server architecture network. All data are stored on a PC disk space and there is no need to store data on the server. The server is used only for files lock and unlock purposes and its disk capacity is used only as an intermediate storage. The method ensures higher information security level than the traditionally used methods. Communication between computers allows to develop applications for cooperative work and documents routing. In a more global aspect the described engine is applicable in both single and multiprogram environments.

Keywords: message transferring engine, network architecture, communicating routines.
1991/95 Mathematics Subject Classification: 94-99.

1. INTRODUCTION

Let us use the term "*server*" to denote the main node of the network, responsible for sharing resources between users. Let us use the term "*client*" to denote workstations on the computer network. Then we would use the terms "*client routine*" and "*server routine*" to denote routines, working on user workstations and server platform. In this way we will describe the typical features of the client-server architecture computer network.

"*Monitor*" denotes a user-written routine, which must take control over user routines (client routines) and is used on client-server architecture network. The method proposed does not use any monitor.

The method requires the server routine to be able to:

— share resources — lock/unlock files or part of their contents;

— store temporary data on server disk storage as intermediate storage.

The operations of sharing the resources are non-interruptible and the server routine cannot affect logically the cross-user communications. Then we will have to discuss only client-routine relationships.

To describe the method, we will use the term "*protocol*" to denote the sequential *actions*, undertaken by the client routines, and the corresponding *states* the client routines may reach during the process of message transfer.

The method serves the following "*protocol*" (the states are shown after the description of every action):

(i) the "*sender*" sends a message to the "*receiver*" ("W");

(ii) the "*receiver*" accepts the *message* sent by the "*sender*", takes some actions and sends an *answer* back to the "*sender*" ("C");

(iii) the "*sender*" accepts the *answer* sent by the "*receiver*" ("M").

It is possible anyone of the two communicating user routines (each of both users) to take the role of the "sender" and consequently the other user routine must take the role of the "receiver".

Because the *answer* does not have only the role of an *acknowledgment*, this protocol is provided to realize the interchange of a wide range of information — every transferred portion of information (message) can have the form of a *request*, *query*, *command*. The type of the message depends on its form. The answer from the receiver can be a *request*, *query*, *command* as a nearer result or a complete document.

The protocol described above facilitates the transfer of any message from the side of the sender and the reply on the side of the receiver after some processing (if needed). Because each of the communicating user routines can take the role of the sender, the protocol is the base for creating channels between the user routines in both the simple and the duplex modes.

2. A METHOD FOR MESSAGE INTERCHANGE

To transfer messages between two users (client routines) it is necessary to:

— establish a connection between the client routines;

— provide resources for the message transfer and sharing.

We will use, whenever it is possible, the term "*user*" instead of "*client routine*", because users communicate through client routines. Every user must identify itself by a user identifier and must place a request for communication in the form of CSB (Communication Sign-on Block), labelled by the same user identifier. Every record in this block must point to the identifier of the other user, with which the CSB

owner wants to communicate. Each client routine must also check the existence of a CSB for each user, marked by a record in its own CSB. A connection between two client routines is considered established when:

(i) CSBs labelled by two user identifiers exist at the same time;

(ii) a record of one CSB points to the corresponding CSB and vice versa.

When (ii) is not satisfied, it means that the CSB points to different (maybe already connected) users.

All resources may be realized as files. In this case the network server must only lock and unlock files (or parts of the files' content) in response to client routines requests.

To provide synchronization between two communicating client routines, a "*post-box*" must be created. In this "post-box", realized as a file too, the information must be recorded as follows:

(i) field "*State*" contains the code of the current state, which the client routines can reach ("W", "C" or "M");

(ii) field "*Identifier*" must point to the identifier of the user, that must receive a message or an answer;

(iii) field "*Message/answer*".

The states shown above are common of the two communicating client routines. It is enough to store only the codes "W", "C" and "M" about the three states, which the user routines can reach. Field "Identifier" points (except when field "State" contains "W") to the user (the client routine), which *must be activated* by the message/answer transfer initiator to receive a message or an answer. To cause changes in the other client-routine state (the client routine to be activated to receive a message or an answer), the client routine, which wants to send a message/answer must:

(i) store into the field "State" a new state code;

(ii) store into the field "Identifier" the user identifier of the corresponding client routine to be activated.

Each client routine must check the "post-box" to determine the state (the states are common of both client routines, as described above).

All read/write operations with a "post-box" information are possible only after the post-box file has been locked. When one of the client routines locks the "post-box" file, the other client routine at this time must wait (or take actions, which must not affect the "post-box" content) until the "post-box" can be locked again. When one of the communicating client routines cannot find its own user identifier stored in the field "Identifier", that routine must immediately unlock the file (except when the file "State" contains the code "W" and the routine wants to send a message). This is possible when there are differences in the speed of the two client routines' execution. In this case immediate unlocking resolves the problem and makes the method independent of the relative speed of the communicating routines' execution.

Now we can describe the protocol used as follows:

(i) the "sender" writes a message to the "receiver" into the field "Message/answer", stores the identifier of the "receiver" in the field "Identifier" and changes the state to "C";

(ii) "receiver" accepts the message sent by the "sender", takes some actions, writes an answer in the field "Message/answer" back to the "sender", stores the identifier of the "sender" in the field "Identifier" and changes the state to "M";

(iii) the "sender" accepts the answer sent by the "receiver" and changes the state to "W".

A client routine must create resources (excluding CSB), e.g. a client routine, executing on the computer of the user with a higher identifier (in lexicographical order). On disconnection, resources must be released in the order of:

(i) disconnection-initiator acting as a "sender" sends the message "quit" to the other client routine (the "receiver");

(ii) the "receiver" releases the corresponding record in its own CSB and returns an answer back to the "sender";

(iii) the "sender" accepts the answer sent by the "receiver" and makes sure that the "receiver" will no more use the shared resources, then the latter releases, in his own CSB, the corresponding record.

Every user (client) routine control flow can be proposed as a graph. The nodes correspond to the states a routine can involve. The arrows are used for signals, which cause changes in the routine state. Fig. 1 displays a graph of two client routines presented by their states and corresponding signals according to

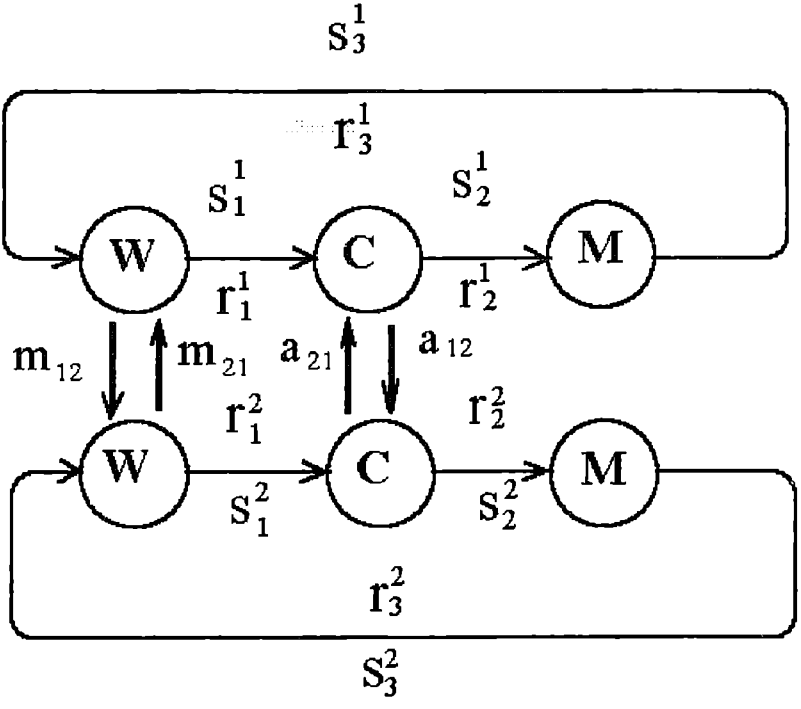


Fig. 1. Graph of communicating user routines

the described method. States "W", "C" and "M" are common to both routines (processes).

Denotifications:

s_1^i: the "sender" sends a message to the "receiver" and changes the state to "C";

s_2^i: the "receiver" changes the state to "M";

s_3^i: the "sender" changes the state to "W";

r_1^j: the "sender" changes the state to "C";

r_2^j: the "receiver" sends an answer to the "sender" and changes the state to "M";

r_3^j: the "sender" changes the state to "W";

m_{ij}: a message from the sender;

a_{ji} : an answer from the "receiver".

3. APPLICATION OF THE METHOD

The method is applied by the author for work with MSDOS — for providing an environment as a remote service facility in the client-server architecture network. To execute MSDOS — a command or a file on a *remote* computer, the command must be transferred to the other client routine, working on the *remote* computer, as a message. To return the messages, produced at the execution time, a file for the messages is used as a temporary disk space on the server. The steps are listed below:

— sending the command, entered by the user, to the client routine, working on the remote computer;

— receiving the message by the remote client routine, executing a command on the remote computer, recording the messages produced in the file for messages and returning an answer to the command initiator via the "post-box";

— receiving the answer, sent by the remote client routine, and typing on the screen the messages, stored into the file for messages.

File-transfer commands are realized as multipartitioned commands, which run on both computers sequentially. Synchronization is provided by the message-transfer method, described above. The steps are listed below:

— To get files from the remote PC disk space:

- in state "W" the command is to be sent to the remote client routine (remote computer);
- in state "C" the remote client routine copies the files into a temporary directory on the server and transfers the answer to the file-transfer initiator;
- in state "M" the initiator receives the answer, displays the messages from the file for messages, and copies the files from the temporary directory on the server into his own disk space.

— To put files into the remote computer disk space:

- in state "W" the initiator copies the files into a temporary directory on the server and transfers the command to the remote client routine;
- in state "C" the remote client routine copies the files from the temporary directory on the server into his own disk space, then sends an answer to the file-transfer initiator;
- in state "M" the initiator receives an answer from the remote client routine and displays the messages from the file for messages.

It is possible for each user to take the initiative for a command execution or file transfer.

4. ADVANTAGES OF THE METHOD

Data security. There is no need to store data for permanent use on the server disk space. Documents can be stored partially on the user's disk space. Each side can access only the part of the information, provided for his (her) own use. High security level is reached when the messages, transferred between the users (the user routines) are encoded. The resource identifiers can be generated as words and/or numbers and they are accessible in the communicating routines at the time of the connection established by them and put in the corresponding GSB records.

Unauthorized access prevention. The user access is protected by a password. Passwords are not registered in the file of common access. The user can change his (her) password. The user must provide secured access to his (her) own data.

Groupwork. The method can be used for transferring data between people, working in a group. In this case the method provides fast access to the information, supported by each of the members of the group.

Coordination. The method provides coordination both between the client routines and the people working in the group.

Documents routing. Documents routing depends on the needs of the organization.

Electronic mail. It is easy to send a message or a document to any group member.

Establishing connections. It is possible to establish connections between every two users. In this way every user can communicate at every time with somebody else. The method provides opportunities for simultaneous connections between the users.

Server machinery requirements. The server can have enough disk space installed, but the latter can be smaller than the space used when the data is stored entirely on the server. There are no special requirements to the server processors.

Workstations requirements. Workstations have usually enough disk space installed. Software can be stored on the server when needed. More space is needed to complete entire documents.

Relationships between applications, based on this method and traditionally written applications. The method, described in this paper, does not exclude the sharing of data files and the application of software on the network server.

5. CONCLUSION

The applications of the method and any further developments are independent of the network and the server types. There is no need to create or use any server monitor routines. All the changes in one application do not affect the other applications. It is not necessary to reconfigure and recompile the server and network software.

The method developed by the author is applicable in cooperative work on computer network, designed on the base of a client-server architecture.

REFERENCES

1. Hoare, C. A. R. Communicating Sequential Processes. Prentice-Hall, London, 1985.
2. Denev, J. Finite State Processes in CSP. In: *Discrete Mathematics and Applications*, Blagoevgrad, 1993.

Received on September 15, 1996
Revised on November 10, 1997

South-West University
Blagoevgrad, Bulgaria

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"

FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 89, 1995

ON THE BRITTLE FRACTURE OF A PIN-JOINTED FRAME

GALJA M. DRAGANOVA, KONSTANTIN Z. MARKOV

The aim of the paper is to report some preliminary results concerning rupture through damage accumulation of a simple pin-jointed frame under tension. Under the elastic and stationary creep conditions (at small strains) this is a well-known problem of strength of material and mathematical theory of creep. Here we assume additionally that damage also evolves in the rods, obeying the classical Kachanov's law, which essentially complicates the problem. In the brittle case, the only one, considered in detail in this paper, the problem is formulated eventually as a coupled nonlinear system of differential equations for the damage variables in the rods. This system, in general, does not admit a close form analytical solution unlike the classical examples of continuum damage mechanics, so that numerical treatment is needed. That is why the special, but realistic case of a common "damage exponent" of the rods is only considered and a simple explicit solution for the damage evolution is found and discussed in more detail.

Keywords: brittle fracture, damage mechanics, pin-jointed frames.
1991 Mathematics Subject Classification: 73M25, 73K05.

1. INTRODUCTION

Consider the pin-jointed frame, shown in Fig. 1. The tensile force $\mathbf{F}$ is applied in the direction of the rod BD. Finding the stresses in such a frame is a well-known exercise in strength of materials, provided the rods behave elastically, see, e.g., [1] and many other textbooks on the subject. If the rods' behaviour is governed by stationary creep law equations, the stresses in the system and, in particular, the

creep rate of the loaded node D, are first found by Kachanov [2], provided the creep deformation is small (so that the so-called elastic analogy applies), see also [3].

Our aim here will be a more detailed investigation of the strain and failure of the frame when damage in the rods appear and evolve following some of the classical schemes of continuum damage mechanics initiated and developed by Kachanov [4, 5], see also [6] for further results and generalizations. Since the rods undergo different stresses, damage within them will reach different levels and will thus lead to a more complicated picture of stress and damage distribution than the ones treated in the classical examples of damage mechanics. In this preliminary stage of our investigation, only the purely brittle case will be dealt with. The problem is rigorously posed in Section 2. But even in this simpler case, unlike the examples of damage mechanics, no simple analytical solution will be possible, since the problem under study will be eventually formulated as a system of two *coupled* nonlinear differential equations governing the damage evolution in the rods which admits, in general, only numerical treatment. That is why the particular, but realistic case of a common "damage exponent" ν of the rods is only considered in Section 3. In this case it appears that the damage parameters of the rods are proportional and a simple explicit solution for the damage accumulation is found in Section 4. This solution is discussed in more detail in the final Section 5.

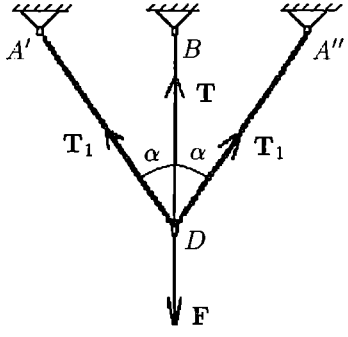


Fig. 1. The pin-jointed frame under study

2. POSING THE PROBLEM

Let all the rods possess in their undamaged state one and the same cross-section S_0 and Young's modulus E^0. Denote as usual by ψ the continuity parameter, so that $\omega = 1 - \psi$ is the damage variable. In the brittle regime under discussion the damage accumulation in a single rod (under uniaxial tension) is governed by the well-known Kachanov's law

$$\dot{\omega} = C \left(\frac{\sigma_0}{\psi} \right)^{\nu}, \tag{2.1}$$

where $\sigma_0 = F/S_0$ is the applied stress, C and ν are material constants [4, 5]. The brittle time-to-rupture, t_b^*, of such a single rod is given then by the known relation

$$t_b^* = \frac{1}{C(1+\nu)\sigma_0^\nu}, \tag{2.2}$$

see again [4, 5]. Hereafter the dimensionless time-scale

$$\tau = t/t_b^* \tag{2.3}$$

will be used, since t_b^* is a natural time-unit for the problem under study.

To derive the damage evolution equations in the rods, let us write down first the only non-trivial statics equation for the problem, namely,

$$2T_1 \cos\alpha + T = F, \tag{2.4}$$

as well as the equation of the compatibility of the strains in the nod D, namely,

$$\varepsilon_1 = \varepsilon \cos^2\alpha. \tag{2.5}$$

Hereafter all quantities with the subscript '1' refer to the rods $A'D$ or $A''D$, and those without a subscript — to the central rod BD. Hence, in particular,

$$\sigma_1 = T_1/S_0, \quad \sigma = T/S_0, \quad \sigma_0 = F/S_0 \tag{2.6}$$

are the stresses in the rods, T_1 and T being the respective magnitudes of the tensile forces in them, see Fig. 1; σ_0 would be the stress in any of the rods if they were single and subjected to the same force F_0. Note also that dealing with brittle fracture solely implies that strains are small, so that the angle α in Eqs. (2.4) and (2.5) remains constant — something that does simplify the study (in the ductile and mixed brittle-ductile failure this angle changes considerably during loading and hence an additional non-linear equation involving this angle should be added to the basic equations).

Assume next that the rods $A'D$ and $A''D$ have the same "damage exponent" ν but different material parameter C_1 in the Kachanov's law (2.1) than the central one BD.[1] This means that Eq. (2.1) applies for the central rod BD, but in the two "side" rods $A'D$ and $A''D$ damage accumulates according to the law

$$\dot{\omega}_1 = C_1 \left(\frac{\sigma_1}{\psi_1}\right)^\nu, \tag{2.7}$$

where $C \neq C_1$. The reason to take different material parameters C and C_1 is that the well-known elementary elastic solution for the frame under study suggests that the central rod is obviously more stressed than the two "side" ones, i.e. $\sigma > \sigma_1$.

[1] Note that the more general case when the exponents ν of the rods differ as well can also be treated without much effort, though no closed form solution is possible. This case will be considered elsewhere.

This means that the central rod will fail faster. That is why, to make the frame more "damage-resistant", one should accordingly choose for the central rod CD a more "damage-resistant" material which accumulates damage slower, i.e. $C < C_1$ at one and the same fixed damage exponent ν. Hence for a given ν and C_1, the dimensionless time-to-rupture of the frame

$$\tau_f^* = t_f^*/t_b^* = \mathcal{T}(C/C_1) \tag{2.8}$$

will be a function of the dimensionless parameter C/C_1, as we shall see below. As a matter of fact, the function $\mathcal{T}$ will be of central importance in our study, since its behaviour (local extrema if any, monotonic decrease and/or increase, etc.) will allow us to draw non-trivial conclusions, concerning optimal "damage-resistant" design of the frame under study, i.e. to get its time-to-rupture τ_f^* as big as possible through an optimal choice of the damage material constants of the rods.

3. BASIC EQUATIONS

Let us now write down the above formulated basic equations in a dimensionless and more convenient form.

First, the equation of statics (2.4) in such a form reads

$$2s_1 \cos\alpha + s = 1\,, \tag{3.1}$$

where

$$s_1 = \sigma_1/\sigma_0\,, \quad s = \sigma/\sigma_0\,, \tag{3.2}$$

with σ_1 and σ_0 defined in Eq. (2.6).

Next, the damage law (2.1) can be recast as

$$\frac{d\omega}{d\tau} = \frac{1}{1+\nu}\left(\frac{s}{\psi}\right)^{\nu}\,, \tag{3.3}$$

see Eqs. (2.2), (2.3) and (3.2). In turn, the appropriate damage law for the side-rods becomes

$$\frac{d\omega_1}{d\tau} = \frac{1}{\xi(1+\nu)}\left(\frac{s_1}{\psi_1}\right)^{\nu}\,, \tag{3.4}$$

where the dimensionless quantity

$$\xi = C/C_1 \tag{3.5}$$

determines, so to say, the relative "damage-resistance" of the central rod as compared to that of the side ones (at a fixed "damage exponent" ν for all rods, let us recall).

To find the stresses in the rods and thus the damage accumulation rates by means of Eqs. (3.3) and (3.4), use is to be made now of the strain compatibility condition (2.5). Recall to this end that the rods are assumed to possess, in the virgin

state ($\psi = \psi_1 = 1$), one and the same Young's modulus E^v. Two possibilities are open now.

First, as the simplest and rough approximation, one can assume that the Young's modulus is not influenced by damage. Then, from Eq. (2.5) (dividing both its sides by $E^v\sigma_0$), one gets

$$s_1 = s\cos^2\alpha\,. \tag{3.6}$$

Together with Eq. (3.1), the latter relation yields the well-known elastic stresses in the rods, namely,

$$s = \frac{\sigma}{\sigma_0} = \frac{1}{1+2\cos^3\alpha}\,, \quad s_1 = \frac{\sigma_1}{\sigma_0} = \frac{\cos^2\alpha}{1+2\cos^3\alpha}\,, \tag{3.7}$$

which therefore are *not affected* by the damage process taking place in the rods. In this way damage accumulation in them is not coupled in the case under study, cf. Eqs. (3.3) and (3.4), and hence they can be solved separately. The failure will have two distinct stages: in the first one all rods will sustain load ($\psi,\ \psi_1 > 0$); in the second stage either the central or the two side rods will already have failed, depending on the ratio ξ, see (3.5), so that the eventual failure will happen when the last of the rods will fail as well. Of course, these two stages will appear in the general case as well, but here, when damage accumulation in the rods is not coupled, the investigation and the appropriate formulae for the time-to-rupture are not difficult to be derived; that is why they will be skipped here.

Instead, let us treat in more detail the more realistic assumption when the current Young's modulus *is influenced* by damage, i.e. $E = E(\psi)$. (This assumption, as well as the idea to measure damage through the observed change in the elastic moduli of a damaging solid, is discussed in detail in [6], where the appropriate references are given as well.) The simplest approximation is to assume that

$$E(\psi) = E^v\psi = E^v(1-\omega) \tag{3.8a}$$

for the central rod and, accordingly,

$$E(\psi_1) = E^v\psi_1 = E^v(1-\omega_1) \tag{3.8b}$$

for the two side-rods, E^v denoting the Young's modulus for the virgin rods. It is noted that such an assumption is natural enough if one recalls the original Kachanov's interpretation of the continuity parameter ψ as the fraction of the undamaged rod cross-section area that only sustains load. Also, this assumption, roughly speaking, reflects the well-known Voigt approximation in mechanics of composite media, if the damage parameter ω is treated, somewhat loosely of course, as the void volume fraction in a porous solid. In this case, noting that

$$\sigma_1 = E^v\psi_1\varepsilon_1\,, \quad \sigma = E^v\psi\varepsilon$$

in virtue of Eqs. (3.8), one finds from Eq. (2.5)

$$\frac{\sigma_1}{\psi_1} = \frac{\sigma}{\psi}\cos^2\alpha \tag{3.9}$$

which, when coupled with Eq. (3.1), yields

$$s = \frac{\psi}{\psi + 2\psi_1 \cos^3 \alpha}, \quad s_1 = \frac{\psi_1 \cos^2 \alpha}{\psi + 2\psi_1 \cos^3 \alpha}. \tag{3.10}$$

Not surprisingly, for undamaged rods ($\psi = \psi_1 = 1$) the purely elastic solution, Eq. (3.7), is recovered once again from Eq. (3.10).

When inserted into Eqs. (3.3) and (3.4), the stresses from Eq. (3.10) now lead to the basic system of coupled differential equations that describes the damage accumulation of the rods, namely,

$$\begin{aligned} \frac{d\psi}{d\tau} &= -f(\psi, \psi_1), \\ \frac{d\psi_1}{d\tau} &= -\frac{1}{A} f(\psi, \psi_1), \end{aligned} \tag{3.11a}$$

with the notations

$$f(\psi, \psi_1) = \frac{1}{1+\nu} \left(\psi + 2\psi_1 \cos^3 \alpha\right)^{-\nu}, \quad A = \frac{\xi}{\cos^{2\nu} \alpha}, \tag{3.11b}$$

since $\omega = 1 - \psi, \omega_1 = 1 - \psi_1$. The system (3.11) should be solved under the natural initial conditions

$$\psi = 1, \quad \psi_1 = 1, \quad \text{at} \quad \tau = 0, \tag{3.12}$$

reflecting the fact that the rods are undamaged at the moment $t = 0$ when loading is applied.

4. SOLUTION OF THE BASIC SYSTEM OF EQUATIONS (3.11)

The solution of the basic initially-value problem (3.11) – (3.12) is elementary. First, dividing equations (3.11a) gives

$$\frac{d\psi}{d\psi_1} = A, \quad \text{i.e.} \quad \psi = A(\psi_1 - 1) + 1,$$

or

$$\omega = 1 - \psi = A\omega_1, \quad \omega_1 = 1 - \psi_1. \tag{4.1}$$

Hence an important consequence of the assumption of common damage exponent ν of the rods is the fact that their damage parameters are proportional, with the proportionality factor A, given in Eq. (3.11b). In this way it turns out that the value of the factor A, i.e. of the dimensionless ratio $\xi = C/C_1$, determines which of the rods will fail first. More precisely:

a) if $A < 1$, i.e. $A = \dfrac{C/C_1}{\cos^{2\nu} \alpha} < 1$ or $C < C_1 \cos^{2\nu} \alpha$, (4.2a)

then the two side-rods fail simultaneously first;

$$\text{b) if } A = 1, \text{ i.e. } A = \frac{C/C_1}{\cos^{2\nu}\alpha} = 1 \text{ or } C = C_1 \cos^{2\nu}\alpha\,, \tag{4.2b}$$

then all rods fail simultaneously;

$$\text{c) if } A > 1, \text{ i.e. } A = \frac{C/C_1}{\cos^{2\nu}\alpha} > 1 \text{ or } C > C_1 \cos^{2\nu}\alpha\,, \tag{4.2c}$$

then the central rod fails first.

It is noted that these results are natural enough since, e.g., the inequality (4.2a) means that the "damage-resistance" of the central rod is considerably higher than that of the side ones because the constant C of the former is considerably less than that of the latter. The central rod accumulates thus damage slower than the other two and, not surprisingly, it ruptures last.

Now, introducing Eq. (4.1) into the second of Eqs. (3.11a) gives

$$\frac{d\psi_1}{d\tau} = -\frac{1}{A(1+\nu)}\left(A' + A''\psi_1\right)^{-\nu} \tag{4.3}$$

with the constants

$$A' = 1 - A\,, \quad A'' = A + 2\cos^3\alpha\,. \tag{4.4}$$

The integration of Eq. (4.3) gives

$$\tau = \frac{A}{A + 2\cos^3\alpha}\left[\left(1 + 2\cos^3\alpha\right)^{\nu+1} - \left(A' + A''\psi_1\right)^{\nu+1}\right]. \tag{4.5}$$

Solving Eq. (4.5) with respect to ψ_1 and using Eq. (4.1) lead to the needed explicit time-dependence of the rods' damage parameters during the loading in the frame under study.

5. DISCUSSION AND CONCLUDING REMARKS

Consider now in more detail the above mentioned three particular cases a) — c), see Eqs. (4.2), in order to determine the eventual time-to-rupture τ_f^* of the frame.

Let first $A < 1$, i.e. the case a) takes place. Then, at the end of the first stage of failure of the frame, when $\psi_1 = 0$ and the side-rods fail, the damage parameter of the central rod has the value $\omega_I = A < 1$, see Eq. (4.1). As it follows from Eq. (4.5), this happens at the moment

$$\tau_I^* = \frac{A}{A + 2\cos^3\alpha}\left[\left(1 + 2\cos^3\alpha\right)^{\nu+1} - \left(1 - A\right)^{\nu+1}\right] \quad (A < 1)\,. \tag{5.1a}$$

In the second failure stage, when $\tau > \tau_I^*$, only the central rod "works", so that one should solve Eq. (2.3) with the initial condition $\omega = \omega_I$ at $\tau = \tau_I^*$ in order to

find the final time-to-rupture, τ_f^*, of the whole frame, corresponding to the moment when $\omega = 1$. Elementary calculations give

$$\tau_f^* = T^{(a)}(\xi) = \tau_I^* + (1 - A)^{\nu+1} \tag{5.1b}$$

with τ_I^* given in Eq. (5.1a). (Note that in this second failure stage $T = F$, so that $s = 1$.)

Let us point out that $\tau_I^* = 0$ at $A = 0$, i.e. $\tau_f^* = 1$ at $A = 0$ as it should be. The reason is that $A = 0$ means that $C_1 = \infty$, so that the two side-rods fail instantaneously and from the very beginning only the central rod sustains load. Moreover, the time-to-rupture τ_f^* as a function of A should be increasing in the interval $A \in [0, 1]$, since increasing A at fixed C and ν (and thus at fixed t_b^*) implies that the parameter C_1 decreases; hence the side-rods become more "damage-resistant" which increases, naturally enough, the life-time of the frame.

Next, from Eqs. (5.1) one immediately finds the time-to rupture τ_f^* in the case b) when all the rods fail simultaneously, just putting $A = 1$ in them:

$$\tau_f^* = T^{(b)}(\xi) = \left(1 + 2\cos^3\alpha\right)^{\nu} \quad (A = 1)\,. \tag{5.2}$$

Let now $A > 1$, i.e. the case c) takes place. Then, at the end of the first stage of failure of the frame, when $\psi = 0$ and the central rod fails, the damage parameter of the side-rods has the value $\omega_I = 1/A < 1$, see Eq. (4.1). This happens at the moment

$$\tau_I^* = \frac{A}{A + 2\cos^3\alpha}\left[\left(1 + 2\cos^3\alpha\right)^{\nu+1} - \left(2\cos^3\alpha\,\frac{A-1}{A}\right)^{\nu+1}\right] \quad (A > 1)\,, \tag{5.3a}$$

as it again follows from Eq. (4.5). In the second failure stage, when $\tau > \tau_I^*$, only the side-rods "work", so that one should solve Eq. (3.4) with the initial condition $\omega = \omega_I$ at $\tau = \tau_I^*$ in order to find the final time-to-rupture, τ_f^*, of the whole frame, corresponding to the moment when $\omega_1 = 1$. Elementary calculations give

$$\tau_f^* = T^{(c)}(\xi) = \tau_I^* + 2^{\nu}(1 - A)^{\nu+1}\cos^{3\nu}\alpha \tag{5.3b}$$

with τ_I^* given this time by Eq. (5.3a). (Note that in this second failure stage $2T\cos\alpha = F$, so that $s_1 = 1/2\cos\alpha$.)

It is noted that $\tau_f^* \to \infty$ at $A \to \infty$, which again is natural. Indeed, at fixed $C > 0$ (in order that the basic time-unit t_b^* makes sense, cf. Eq. (2.2)) $A \to \infty$ only if $C_1 \to 0$, so that in the limit $A = \infty$ the side-rods do not accumulate damage. The only damage phenomenon will be in this case the failure of the central rod which will happen at the moment

$$\lim_{A\to\infty} \tau_f^* = \left(1 + 2\cos^3\alpha\right)^{\nu+1} - 2^{\nu+1}\cos^{3(\nu+1)}\alpha\,,$$

as it follows from Eq. (5.3a).

Combining now the formulae (5.1) to (5.3) gives

$$\tau_f^* = \mathcal{T}(\xi) = \begin{cases} \mathcal{T}^{(a)}(\xi), & \text{if } \xi < \cos^{2\nu}\alpha, \\ \mathcal{T}^{(b)}(\xi), & \text{if } \xi = \cos^{2\nu}\alpha, \\ \mathcal{T}^{(c)}(\xi), & \text{if } \xi > \cos^{2\nu}\alpha, \end{cases} \tag{5.4}$$

which accomplishes our aim — analytic evaluation of the function $\mathcal{T}(\xi)$, see Eq. (2.8), that gives the time-to-rupture τ_f^* of the whole frame for a given dimensionless ratio $\xi = C/C_1$ of Kachanov's material parameters of the rods (with a fixed and common "damage exponent" ν). The superscripts in Eq. (5.4) correspond obviously to the three different situations a) — c) of frame failure, discussed in Section 4, see Eq. (4.2).

For illustration the plot of the function $\tau_f^* = \mathcal{T}(\xi)$ for a typical angle $\alpha = \pi/4$ and $\nu = 3$ is shown in Fig. 2.

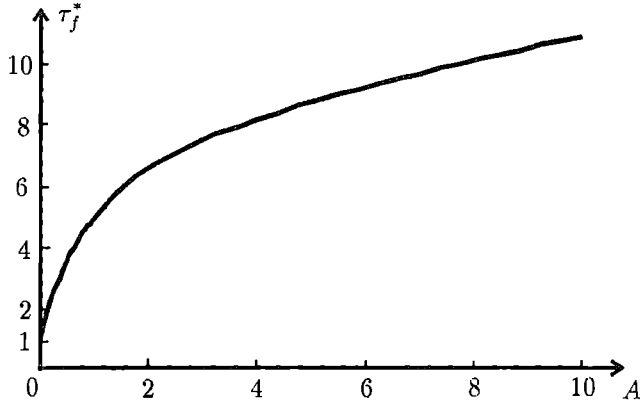


Fig. 2. Dimensionless time-to-rupture τ_f^* of the frame as a function of the parameter $A = (C/C_1)/\cos^{2\nu}\alpha$ at $\alpha = \pi/4$ and $\nu = 3$

A more detailed numerical investigation shows that τ_f^* is always a monotonically increasing function of the ratio C/C_1. This means in the damage mechanics context that for a given central rod one should add side-rods for which C_1 is as small as possible, i.e. their "damage-resistance" is as high as possible. Of course, this result should have been expected qualitatively. The above analysis allows us, however, to draw quantitative conclusions as well, i.e. to evaluate simply the relative time-to-rupture increase of the frame as compared to that of the central rod if it were a single one and subjected to the same tensile force **F**.

ACKNOWLEDGEMENTS. The support of this work by the Bulgarian Ministry of Education, Science and Technology under Grant No MM416-94 is gratefully acknowledged.

REFERENCES

1. Gastev, V. A. A brief course on strength of materials. 2nd ed., Nauka, Moscow, 1977 (in Russian).
2. Kachanov, L. M. Theory of creep. Gos. Izd. Fiz.-mat. Lit., Moscow, 1977 (in Russian).
3. Odqvist, F. K. G. Mathematical theory of creep and creep rupture. Oxford University Press, 1974.
4. Kachanov, L. M. T On the time to rupture in creep. *Izv. AN SSSR, Otd. Tehn. Nauk*, **8**, 1958, 26–31 (in Russian).
5. Kachanov, L. M. Introduction to continuum damage Mechanics. Klüwer Acad. Publ., 1986.
6. Lemaitre, J., J.-L. Chaboche. Mécanique des Materiaux Solids. Dunod, 2-ème ed., Paris, 1988.

Received on July 17, 1996
Revised on October 21, 1997

Galja M. Draganova
Institute of Chemical Technology
and Biotechnology
3 blvd. Aprilsko vastanie
POB 110, BG-7200 Razgrad
Bulgaria

Konstantin Z. Markov
Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5 blvd J. Bourchier
BG-1164 Sofia, Bulgaria
e-mail: kmarkov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 89, 1995

SOME EXAMPLES OF LEXICOGRAPHIC ORDER ALGORITHMS AND SOME OPEN COMBINATORIAL PROBLEMS

DIMITAR L. VANDEV

A general reasoning based on the lexicographic order is studied. It helps to create algorithms for generation of sets of words having certain natural and good properties. Several examples are considered and the performance of the proposed algorithms is calculated. An open combinatorial problem regarding the set of partitions arises.

Keywords: enumerating algorithms, lexicographic order functions.
1991/95 Mathematics Subject Classification: 68E05, 65C20.

1. INTRODUCTION

There are numerous examples of sets of words — vectors of natural numbers, which as one set of entities may be used for some computational purposes: sets of all permutations, combinations and many others. In many cases one needs to go across such a set and perform some computations for each member. (See [3] for many examples in combinatorial calculations.)

The problem of the efficient generation of all elements of a class of combinatorial configurations with given properties is considered as an important problem in the theory of algorithms. The generation in a prescribed lexicographical order is one of the most investigated cases, see [5, 4].

In the present paper an attempt is made to use the lexicographic order of these words as a tool for creating enumerating (or generating) algorithms. It turns out that the proposed scheme is useful also for calculating the performance of the

algorithms. In some cases it is possible to calculate it easily, while in others an open problem arises.

A part of these examples were presented as a short communication at the Seminar of Statistical Data Analysis in Varna, Bulgaria, see [7].

2. DEFINITIONS AND NOTATIONS

Let N be the set of natural numbers $\{0, 1, \ldots, n\}$. Call N alphabet. Denote by $S = S(m, n)$ the m-times Cartesian product of the set N. The elements of the set S are called words with a fixed length m and a common alphabet N. It is clear what a lexicographic order in this set means. One word is called "larger" than another if its first (after the common beginning of the two words) letter is larger than the corresponding letter of the second word. Note that the set of numbers (with leading zeroes) with digits from N is ordered in the same manner.

For any subset W of the set S this order induces the same order for the elements of W. To make things more formal, we shall call the word formed by the first k letters of the word w prefix and denote it by $w(k)$. The notation $w(l, k)$, $l \leq k$, will be used to denote the set of letters in the places from l to k. For prefixes with fixed length we have the same induced from S linear order. Formally, the empty word $w(0)$ is the unique element of the set $S(0, n)$. We introduce two sets of mappings (projections preserving the order) from S onto W. If the word belongs to W, these mappings will preserve its prefix. In the following we shall consider the set W fixed.

Definition 1. For $s \in S$, $z = \mathbf{First}(k, s)$ is the first member $z \in W$, such that $z(k) \geq s(k)$, if it exists. In any other case it is the first member of W.

Definition 2. For $s \in S$, $z = \mathbf{Last}(k, s)$ is the last member $z \in W$, such that $z(k) \leq s(k)$, if it exists. In any other case it is the last member of W.

If $w \in W$, $z = \mathbf{First}(k, w)$ is the first member of W and $z = \mathbf{Last}(k, w)$ is the last member of W with the same prefix $z(k) = w(k)$. So the element $w_0 = \mathbf{First}(0, w)$ is the very first in W and $l_0 = \mathbf{Last}(0, w)$ — the very last in the global linear order.

We shall introduce also a mapping $\mathbf{Increase}(k, w)$, which will be used to increase only the k-th letter of the word w. This mapping is not defined for all elements of W or S. Moreover, its result (when defined) is not obliged to belong to the set W.

Definition 3. We say that $\mathbf{Increase}(k, w)$ equals the smallest word $z \in S$, such that $w(k) < z(k)$, if this word exists. In any other case it is not defined.

Obviously, not for all elements of W this definition will lead to increasing of only and exactly the k-th letter.

3. MAIN RESULTS

First we shall prove some simple consequences of these natural definitions. Then an algorithm will be presented and a theorem about its completeness will be proved. Then a simple theorem which helps to estimate the effectivity of the

algorithm will be stated. It will concern the mean number of steps needed to produce the next to w word in W.

Lemma 1. *Both mappings* **First** *and* **Last** *are well defined and idempotent with fixed* k: $\mathbf{First}(k, \mathbf{First}(k, s)) = \mathbf{First}(k, s)$, $\forall s \in S$.

Proof. Denote by $l_0 = \mathbf{Last}(0, s)$. Then the set S may be split into two subsets: $S = S_1 + S_2 = \{s : s(k) \leq l_0(k)\} + \{s : s(k) > l_0(k)\}$.

When $s \in S_1$, the image $z = \mathbf{First}(k, s)$ exists, because it may be represented as intersection of non-empty subsets of W. It is unique because of the linear order. When $s \in S_2$, we have $w_0 = \mathbf{First}(k, s)$ according to the definition. The second statement is obvious because of the definition too. The same argument works for the mapping **Last**. □

Lemma 2. *If* $z = \mathbf{Last}(k, w)$, *then for each* $i \geq k$, $z = \mathbf{Last}(i, z)$. *If* $z = \mathbf{First}(k, w)$, *then for each* $i \geq k$, $z = \mathbf{First}(i, z)$.

Proof. Suppose that $z = \mathbf{Last}(k, w)$, but $z < \mathbf{Last}(i, z)$ for $i > k$. We have $w \leq z = \mathbf{Last}(k, z)$. Then $\mathbf{Last}(k, z) < \mathbf{Last}(i, z)$. However, the first k letters of these two words coincide — which is a contradiction. The same reasoning works for the dual statement. □

Let fix $w \in W$ and consider the set of equations $w = \mathbf{Last}(k, w)$. Note that $w = \mathbf{Last}(m, w)$ is true for each w. According to Lemma 2, there exists a minimal k for which this equality holds. For the last word in W we shall have $k = 0$. Again the same is true for the first word in W (in this case the mapping **First** should take place in the equations).

These considerations give us the possibility to construct the following algorithm for consecutive computing of the 'next' to w word in the set W:

```
function next(word)
1  k = n;
2       while word = Last(k, word);
3            k = k − 1;
4       end_while;
5            if k = 0 stop;
6       word := Increase(k, word);
7       word := First(k, word);
end
```

This algorithm needs some explanations. Lines 1–4 perform the search for the largest k for which $w \neq \mathbf{Last}(k, w)$. As the purpose of these lines is to find the integer k, it seems natural to combine them into a function: $k = \mathbf{Last\_Not\_Last}(w)$. Another reason to make this will be seen in the examples — in most cases it is easier to calculate the function **Last_Not_Last** than the mapping **Last**. The number k may be easily interpreted as the position of the first letter changing when moving from the given word to the next one in a lexicographically ordered set W. Line 5 prevents the use of the program after the last word $l_0 = \mathbf{Last}(0, w)$ has been reached. Line 6 increases the k-th letter of the word to the next letter allowed

(given the prefix $w(k-1)$). Line 7 simply uses the mapping **First**, but the prefix is now one letter longer — $w(k)$.

Usage of the function **Last_Not_Last** simplifies the algorithm:

```
function next(word)
  k = Last_Not_Last(word);
      If k = 0 stop;
  word := Increase(k, word));
  word := First(k, word);
end
```

Theorem 1. *Starting with* $w_0 = \mathbf{First}(0, w)$, *the above algorithm exhausts all elements of the set* W, *i.e.* $l_0 = \mathbf{Last}(0, w)$ *is reached.*

Proof. The first thing is to check the possibility to define and use the mapping **Increase** properly. Let $k > 0$. As the element w is not equal to $\mathbf{Last}(k, w)$, there exists a word $z \in W$, such that $z(k) = w(k)$ and $w < z$. Let us choose the next to w element $z \in W$. Suppose now that the k–th elements of z and w are equal. This means that $z(k) = w(k)$ and we have $w < z$, but $\mathbf{Last}(k, w) = w$. This is a contradiction. Thus, there exists a word $z \in W \subset S$ with greater k-th letter. Such a word exists in S. So in our algorithm we may use the function **Increase** when the proper $k > 0$ is found.

Now we shall show that no word $z \in W$ may be skipped by the algorithm. If $m = 1$, the statement follows from the definitions of **Increase** and **First**. If **Increase** does not produce a word from W, then this will be done by **First**.

The induction on m uses the fact that each part of the set W with a fixed prefix uses the same definitions of the functions **First**, **Last**, **Increase**. If for any fixed first letter the algorithm is exhaustive, it will be exhaustive for the whole set. □

Comment 1. The study of this simple proof shows that the definition of **Increase** may be made more complicated — not simply to increase the corresponding letter, but to choose it in such a way that the corresponding prefix "belongs" to W. The function **First** does not need to be defined for any word in S. For the index k achieved at the first step, there always exists a number in the alphabet put at the k-th place in the word, so that $\mathbf{First}(k, w) := \mathbf{First}(z, w)$ is well defined. This situation is effectively explored in some of the examples below.

Comment 2. It is easy to see that if one defines the mapping **Increase** to do nothing when $k = 0$, the proposed algorithm will loop infinitely across the set W starting from the beginning again and again.

Comment 3. The same argument may be used for the statement concerning the reverse order. The mapping **First** may be replaced by **Last**, the mapping **Increase** — by the correspondingly defined mapping *Decrease*. All the statements above will remain true except for the order — it will become the inverse order. There is one more formal duality in the lexicographical order – the interchanging of the letters. The most natural interchanging is to read the word backward. Then

the first letters of the word are changed while the last are kept fixed. We shall call such an order dual. With any set four different orderings into the set S can be defined. The definitions above are to be changed correspondingly for any such order. Any of these orderings may be useful to consider when an enumeration is performed.

Theorem 2. *Denote by W_k the set of all different prefixes $w(k)$ of words in W. We shall assume that $|W_0| = 1$ and $W_n = W$. Suppose that for each $k = 0, 1, 2, \ldots, n-1$, we have $|W_k|/|W_{k+1}| < q < 1$. Then according to the uniform distribution on W the expected number of steps to reach the address of change $\mathbf{E}(n-k)$ fulfills the inequality*

$$\mathbf{E}(n-k) < \frac{2}{1-q}.$$

Proof. The set W may be represented as a graph — a tree with totally $N+1$ vertices and N edges. Each vertex corresponds to a fixed prefix. Then the total way of our algorithm to generate all the members of the set is proportional to $2N$. Denote $r_k = |W_k|$. The expected number $\mathbf{E}(n-k)$ may be represented in the form

$$\mathbf{E}(n-k) \le \frac{2N}{|W|} \le \frac{2}{r_n}(r_0 + r_1 + r_2 + \cdots + r_n)$$

$$= 2\left(\frac{r_0 r_1 \ldots r_{n-1}}{r_1 r_2 \ldots r_n} + \frac{r_1 r_2 \ldots r_{n-1}}{r_2 r_3 \ldots r_n} + \cdots + 1\right) \le 2\frac{1-q^{n+1}}{1-q}.$$

Comment 4. It is clear that the assumption of the theorem may be weakened in a number of ways. For example, it is enough for k to run over the set $0, 1, \ldots, n-j$. Then the expectation will be limited by $j + 1/(1-q)$.

As we shall see in the next section, despite that the assumption is not fulfilled in many cases, the average number of steps remains finite. On the other hand, the set consisting of two words $\{1, 1, 1, \ldots, 1\}$ and $\{2, 1, 1, \ldots, 1\}$ will need an expected number of steps proportional to n. It is an open question, what, in general, happens to the expected number of steps when all the dualities mentioned in Comment 3 are explored.

4. EXAMPLES

In the next examples we shall construct the mappings **First**, **Last**, *Increase* and the function **Last_Not_Last** for different subsets of S. We shall try also to calculate the computational complexity of the algorithms. In fact, one needs only the distribution of k — "the place of first change" in lexicographically ordered words. It is clear that the proposed algorithm will be as effective as closer to n the expectation of this place is situated. For that purpose one has to calculate also the size of the corresponding data set and assume uniform probability on it. So, the mean effort for constructing the next element should represent the complexity of the embedding of the data set in the given order. It will be seen that the use of different embeddings is of primary interest.

The first two examples have been extensively studied in [4, 5]. Here they are mentioned only to show that the idea we use leads to natural and well-known algorithms.

4.1. PERMUTATIONS OF N ELEMENTS

In Table 1 a part of the set of all permutations of 5 elements is shown in a lexicographic order.

TABLE 1. Part of permutations of 5 elements

1 2 3 4 5	1 3 2 4 5	1 4 2 3 5	...
1 2 3 5 4	1 3 2 5 4	1 4 2 5 3	...
1 2 4 3 5	1 3 4 2 5	1 4 3 2 5	...
1 2 4 5 3	1 3 4 5 2	1 4 3 5 2	...
1 2 5 3 4	1 3 5 2 4	1 4 5 2 3	...
1 2 5 4 3	1 3 5 4 2	1 4 5 3 2	...

It is clear that the mapping **Last** simply orders all the elements of w after (and including) the k-th one in a decreasing order, while for the mapping **First** this order is increasing.

The function **Last_Not_Last** finds the smallest k such that after it all elements are in a decreasing order. Denote $j = n - k$. Then it is clear that j runs from 1 to n. For a given k, this function needs j subtractions and comparisons.

The mapping **Increase** is more complicated. It takes the next larger than $w(k,k)$ integer from the set of integers $w(k,n)$ and should replace it with $w(k,k)$. The last step **First** is equivalent to inversion of the sequence of the last j integers.

Theorem 2 can be applied to the set of permutations with k running up to $n-j$ and $q = 1/j!$. However, it might be interesting to calculate exactly the expected number of steps of the algorithm. This is done in [5, Section 5.1], in the terms of transpositions and comparisons. The expected number of integer calculations then is proportional to $(e-1)$ and remains finite as $n \to \infty$.

4.2. SUBSETS OF M ELEMENTS OUT OF SET OF N

In Table 2 the set of all subsets of 4 elements, taken from the set of 6 elements, is shown in a lexicographic order. One calls the objects combinations of n elements of class 4. Here the letters are kept in an increasing order inside the word — they should not coincide.

TABLE 2. Subsets of 4 elements out of 6

1 2 3 4	1 3 4 5	2 3 4 6
1 2 3 5	1 3 4 6	2 3 5 6
1 2 3 6	1 3 5 6	2 4 5 6
1 2 4 5	1 4 5 6	3 4 5 6
1 2 4 6	2 3 4 5	

The mapping **Last** changes the last $m-k$ elements of w into the largest elements of N, while **First** sets these elements to the smaller ones following the k-th element of w. The function **Last_Not_Last** finds the largest k such that $w(k,k)$ is not equal to $n-(m-k)$. The mapping **Increase** simply adds 1 to the corresponding element of w. This is also well-known algorithm [5, Section 5.2.2].

The number of combinations of n elements class m is $\binom{n}{m}$. In a similar way, as in permutations, we find the number of calculations as a function of $j=m-k$ to be about $4j$. The distribution of j is also easy to construct:

$$p_j = \frac{\binom{n-m-1-j}{j}}{\binom{n}{m}}.$$

So, we come to even stronger result, namely, that with the growth of $n-m$ the expected number of calculations decreases.

Here the application of Theorem 2 is also possible, which yields $|W_k| = \binom{n-(m-k)}{k}$

4.3. PARTITIONS OF AN INTEGER I

To generate the set W of all partitions of a given integer n into a sum of number integers is an easy problem for this algorithm. In Table 3 all partitions of 10 into the sum of up to 4 numbers are given (except the trivial 000 10). This presentation allows to split easily W into partitions of exactly 2, 3 and 4 non-zero numbers. These subsets follow consecutively. In the next examples other representations will be used.

TABLE 3. Partitions of 10 into up to 4 members

0 0 1 9	0 1 3 6	1 1 2 6
0 0 2 8	0 1 4 5	1 1 3 5
0 0 3 7	0 2 2 6	1 1 4 4
0 0 4 6	0 2 3 5	1 2 2 5
0 0 5 5	0 2 4 4	1 2 3 3
0 1 1 8	0 3 3 4	2 2 2 4
0 1 2 7	1 1 1 7	2 2 3 3

The mapping **Last** distributes the remaining portion of n into maximum equal portions among the remaining numbers after the k-th one. The mapping **First** states all these numbers to $w(k,k)$ and the remainder from m is added to the last number. The function **Last_Not_Last** finds the largest k such that $w(m,m) - w(k,k) \geq 2$. The mapping **Increase** simply increases by 1 the corresponding element of w. This algorithm is due to Hindenburg (see [1, Section 14.3]).

In order to apply the theorem, we have to calculate $w_k = |W_k|$. This is not an easy problem, however. Consider the unlimited case — all partitions of n will be fixed as words of length n with non decreasing elements.

One sees that the number w_k is a sum of partition numbers, subjected to two kinds of restrictions — concerning the maximum number of elements and the size of the largest element.

We have $w_0 = 1$ and $w_k = |w_{k-1}| + 1$ until $n - k \geq [n/2]$. Starting from $[n/2] + 1$ until $n - k = [n/3]$, we have $w_k = w_{k-1} + 2$ or $w_k = w_{k-1} + 3$. Here the choice depends on the remainder of the division on 3.

This example shows that in this case Theorem 2 is not applicable. Indeed, $w_{[n/2]+1}/w_{[n/2]} = 1 + 1/n$ and it tends to one as n increases. Nevertheless, we hope that the average number of steps is finite in this representation also. The exact statement remains an open problem.

4.4. PARTITIONS OF AN INTEGER II

Here we use the representation which follows from the formula

$$\sum_{i=1}^{n} i n_i = n.$$

Here the numbers n_i represent the number of members of size i in a particular partition. The number of members in each partition is $\sum_{i=1}^{n} n_i$. It is clear how to convert one representation into another. Table 4 contains the set of all partitions of 10 into members less than 8, but in another lexicographical order according to the new presentation. Instead of the restriction on the number of members, we now pose a restriction on the maximal member of the partition.

Here we shall exploit the dual order and present the same partitions in an order with a fixed suffix. In this order it is easy to add the additional restriction on the maximum member of the partition, say, it is equal to 7: starting with $a_1 = N$ this new algorithm produces all partitions of 10 to members not bigger than 7.

TABLE 4. All partitions of 10 into numbers up to 7

10 0 0 0 0 0 0	4 0 2 0 0 0 0	0 0 2 1 0 0 0	4 0 0 0 0 1 0
8 1 0 0 0 0 0	2 1 2 0 0 0 0	2 0 0 2 0 0 0	2 1 0 0 0 1 0
6 2 0 0 0 0 0	0 2 2 0 0 0 0	0 1 0 2 0 0 0	0 2 0 0 0 1 0
4 3 0 0 0 0 0	1 0 3 0 0 0 0	5 0 0 0 1 0 0	1 0 1 0 0 1 0
2 4 0 0 0 0 0	6 0 0 1 0 0 0	3 1 0 0 1 0 0	0 0 0 1 0 1 0
0 5 0 0 0 0 0	4 1 0 1 0 0 0	1 2 0 0 1 0 0	3 0 0 0 0 0 1
7 0 1 0 0 0 0	2 2 0 1 0 0 0	2 0 1 0 1 0 0	1 1 0 0 0 0 1
5 1 1 0 0 0 0	0 3 0 1 0 0 0	0 1 1 0 1 0 0	0 0 1 0 0 0 1
3 2 1 0 0 0 0	3 0 1 1 0 0 0	1 0 0 1 1 0 0	
1 3 1 0 0 0 0	1 1 1 1 0 0 0	0 0 0 0 2 0 0	

The function **Last_Not_Last** finds first k from the beginning such that $k \leq l = \sum_{i=1}^{k-1} i.a_i$. Then a_k is increased by one. **First** nullifies all elements in the beginning and makes $a_1 = l - k$.

Let us try to estimate the performance of the algorithm. As usual $P(n) = |W|$ is the number of partitions of n. Denote by r_k the number of different suffixes in the words of the set W. It is clear that $r_{n-1} = P(n), r_0 = 1, r_1 = 2$.

Consider now the case: $3 \leq k \leq n-1$. It is clear that the set R_k may be mapped into the set R_{k+1}, so that the additional element (at place $m = n - k$) is zero. However, it is possible to make one more mapping of the same set R_k into elements of R_{k+1} with non-zero elements at the same place m. This can be done if every suffix is shifted left until the first non-zero element occupies the place m. The remaining portion of the suffix is filled with zeros. The only exception for this second mapping is the suffix consisting of zeros only. So we have the inequality

$$r_k/r_{k+1} \leq 1/2 + 1/r_{k+1} \leq 5/6.$$

According to Theorem 2 this means that the mean number of steps is finite and does not increase as $n \to \infty$. The exact value of this mean, as well as the exact distribution of the number of steps remain an open problem in this presentation also.

4.5. BELL POLYNOMIALS

For exact definitions see [3, Ch. I, p. 10]. This example is included merely to illustrate the use of the algorithm working in the reverse order. Consider the set of all vectors of natural numbers satisfying the following two equalities:

$$\sum_{i=1}^{\infty} k_i = m, \qquad \sum_{i=1}^{\infty} i k_i = n.$$

The summation above is assumed to be infinite for simplicity. It is clear that only the first $n - k + 1$ elements may be non-zero. This set is of some interest in many applications. In addition to computing Bell polynomials, it is used in the distributions of order k. Again we have partitions and the problem could be solved using the first representation of partitions in Section 4.3 and then screening partitions with number of members less than m. However, we shall give here an explicit solution.

In Table 5 the solutions for $n = 13$ and $m = 6$ are shown in reverse order. The reverse order is chosen because of the simplicity of the mapping **Last** in this case.

TABLE 5. Bell Polynomials $n = 13$ and $m = 6$

5 0 0 0 0 0 0 1	3 0 2 1 0 0 0 0
4 1 0 0 0 0 1 0	2 3 0 0 1 0 0 0
4 0 1 0 0 1 0 0	2 2 1 1 0 0 0 0
4 0 0 1 1 0 0 0	2 1 3 0 0 0 0 0
3 2 0 0 0 1 0 0	1 4 0 1 0 0 0 0
3 1 1 0 1 0 0 0	1 3 2 0 0 0 0 0
3 1 0 2 0 0 0 0	0 5 1 0 0 0 0 0

The reverse algorithm will be used. The mapping *Decrease* decreases by 1 the corresponding element of w. The mapping **Last** sets all elements with indexes

greater than k to zero, then $w(k+1,k+1) = m_k - 1$, and finally adds 1 to the number in position $k+n_k-(k+1)*m_k$. Here m_k and n_k are the values of the sums in the definition of Bell polynomials, taken over indexes greater than k. The function *First_Not_*__First__ finds the largest k such that $w(k,k) > 0$ and $w(k,k) \leq max - 2$. Here max is the index of rightmost non-zero element.

In order to estimate the performance of the algorithm in this situation, we shall point out that the representation of partitions given in Section 4.3 is much more economical. Moreover, the additional restriction $m = \sum_{i=1}^{\infty} k_i$ in this case is in concordance with the presentation. It means that it makes no sense to use this second representation if one needs effectivity.

4.6. PARTITIONS WITH AN ADDITIONAL RESTRICTION [1]

Let us consider the algorithm for the following partition problem with the additional restriction:

$$\sum_{i=1}^{l} k_i = m, \qquad \sum_{i=1}^{l} {k_i}^2 = n.$$

By using the algorithm described in Section 4.3 and simply screening the second equation, an easy solution could be given of this problem. As an example, the results are presented in Table 6 for words of length 15, $m = 16$ and several values of n.

TABLE 6. $\sum_{i=1}^{15} k_i = 16, \sum_{i=1}^{15} {k_i}^2 = n$

n	
18	1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
20	0 1 1 1 1 1 1 1 1 1 1 1 1 2 2
22	0 0 1 1 1 1 1 1 1 1 1 1 2 2 2
	0 1 1 1 1 1 1 1 1 1 1 1 1 1 3
24	0 0 0 1 1 1 1 1 1 1 1 2 2 2 2
	0 0 1 1 1 1 1 1 1 1 1 1 1 2 3
26	0 0 0 0 1 1 1 1 1 1 2 2 2 2 2
	0 0 0 1 1 1 1 1 1 1 1 1 2 2 3
28	0 0 0 0 0 1 1 1 1 2 2 2 2 2 2
	0 0 0 0 1 1 1 1 1 1 1 2 2 2 3
	0 0 0 1 1 1 1 1 1 1 1 1 1 3 3
	0 0 1 1 1 1 1 1 1 1 1 1 1 1 4

It would be interesting to investigate the combinatorial properties of this set and to study the properties of the algorithm in this case.

[1] The author is grateful to Prof. G. Zbaganu who mentioned this problem at the 8-th Seminar on Statistical Data Analysis, Varna, 1992, and then helped to reformulate the algorithm.

Consider the set $F_n^{(m)}$ of words of fixed length n, consisting of zeroes and ones, and having the property that they do not contain m or more than m consecutive ones. (This example was proposed by P. Mateev.)

It may be easily proved that the cardinalities $f_n^{(m)}$ of $F_n^{(m)}$ satisfy the following recurrent relation:

$$f_n^{(m)} = \sum_{i=1}^{m} f_{n-i}^{(m)}. \tag{4.1}$$

When $m = 2$, these numbers form the well-known Fibonacci sequence. For arbitrary m and starting conditions $f_0^{(m)} = 0$ and $f_1^{(m)} = 1$, Gabai [2] called them Fibonacci numbers of order M. Philippou [6] calculated them as sums of multinomial coefficients. For the words of zeroes and ones, however, the starting conditions are $f_0^{(m)} = 1, f_1^{(m)} = 1$, as with the original Fibonacci numbers. For the particular case $m = 3$ we have $f_n^{(3)} = 1, 1, 2, 4, 7, 13, 24, \ldots; n = 0, 1, 2, \ldots$ In Table 7 all the $f_5^{(3)}$ zero-one words are presented.

TABLE 7. Fibonacci words for $n = 5$ and $m = 3$

0 0 0 0 0	0 1 0 0 1	1 0 0 1 1
0 0 0 0 1	0 1 0 1 0	1 0 1 0 0
0 0 0 1 0	0 1 0 1 1	1 0 1 0 1
0 0 0 1 1	0 1 1 0 0	1 0 1 1 0
0 0 1 0 0	0 1 1 0 1	1 1 0 0 0
0 0 1 0 1	1 0 0 0 0	1 1 0 0 1
0 0 1 1 0	1 0 0 0 1	1 1 0 1 0
0 1 0 0 0	1 0 0 1 0	1 1 0 1 1

The algorithm for generating such n-tuples is extremely simple. The function **Last_Not_Last** has to find the first zero preceded by less than $m-1$ ones, **Increase** puts one at this place and **First** nullifies all elements with greater indexes.

Here the mean value of the needed calculations is obviously proportional to the place $j = n - k$ of the zero to change. Denote this mean value by j_n. The above recurrence relation then leads to a new relation for the mean values. In order to obtain this relation, we shall use the proof of the recurrence formula (4.1). All n-words may be divided into M disjoint subsets $S_1, S_2, \ldots, S_m$ (we suppose that n is large enough). The l-th subset S_l has an arbitrary prefix and last numbers are fixed:

$$S_l = \{w \in F_n^{(m)} : (w_1, w_2, \ldots, w_{n-l}, 0, 1, 1, 1, \ldots, 1)\}.$$

These subsets cover the whole set $F_n^{(m)}$. The cardinalities of the sets are clearly $f_{n-l}^{(m)}$. In each subset the algorithm stops at the first 0, performing exactly l steps, or enters the prefix looking for the next available zero. In the first case the prefix

should end with exactly m numbers — a zero and $m-1$ ones. Its cardinality equals to $f_{n-l-m}^{(m)}$. So we come to the formula

$$j_n = \frac{\sum_{i=1}^{m}((j_{n-i}+i)(f_{n-i}^{(m)} - f_{n-i-m}^{(m)}) + if_{n-i-m}^{(m)})}{\sum_{i=1}^{m} f_{n-i}^{(m)}}$$

$$= \sum_{i=1}^{m} w_i(j_{n-i} - j_{n-m-i}) \quad + \sum_{i=1}^{m} iw_i.$$

The proportions $w_l = f_{n-l}^{(m)}/f_n^{(m)}$, $l = 1, 2, \ldots, m$, are fixed as $n \to \infty$. The sequence j_n then obviously converges to the finite number $\sum_{i=1}^{m} iw_i * \lim(f_{n+m}/f_n)$.

5. PERFORMANCE AND CONCLUSION

Both forms of the algorithm have quite different performances. For the first it is quadratic in j. One hardly expects a comparison of two words of length j to be made for shorter time. It may be expected a good performance from the second form when the function **Last_Not_Last**, as well as the mapping **First** depend linearly of j. In all examples above this was made possible.

In the case when the distribution of j has a finite mean, not depending of n, the asymptotic properties of the algorithm are extremely nice.

We do not know the distribution of j in the case of Bell polynomials and the performance of the presented algorithms in this case remains an open problem which would be interesting to be solved.

It is clear that building up programs in such a way, one can hardly expect that they will be fast without some additional efforts. However, in all cases above it turned out that only slight modifications were needed to make the programs work quite satisfactorily.

Another useful hint may be to try the other orders to change the mappings **First**, **Last** and **Increase**, correspondingly, and to see what will happen to the program. It may became shorter and faster.

Acknowledgements. The author is grateful to K. Manev, whose useful comments have improved the presentation. The work was partially supported by the Bulgarian Ministry of Education, Science and Technology under Grant No MM440/94.

REFERENCES

1. Andrews, G. E. The theory of partitions. Encyclopedia of Mathematics and its Applications, vol. 2, Addison-Wesley, Reading, Massachusetts, 1976.
2. Gabai, H. Generalized Fibonacci k-sequences. *Fibonacci Quarterly*, **8**(1), 1970, 31–38.

3. Kaufmann, A. Introduction á la Combinatorique en Vue des Applications. Dunod, Paris, 1968.
4. Lipski, W. Kombinatorics for Programists. Wydwnictwa Naukowo-Techniczne, Warszawa, 1982 (in Russian).
5. Reingold, E. M., J. Nivergeld, N. Deo. Combinatorial Algorithm. Theory and Practice. Prentice-Hall, Englewood Cliffs, 1977.
6. Philippou, A. N. A note on the Fibonacci sequence of order k and multinomial coefficients. *Fibonacci Quarterly*, **21**(2), 1983, 82–86.
7. Vandev, D. L. Alphabetical ordering — useful tool for creating algorithms. In: *Statistical Data Analysis — SDA92, Proceedings*, Seminar on Statistical Data Analysis, Varna, Bulgaria, September 1992, 121–126.

Received on September 23, 1996
Revised on November 5, 1997

Faculty of Mathematics and Informatics
"St. Kl. Ohridski" University of Sofia
5 blvd J. Bourchier
BG-1164 Sofia, Bulgaria
e-mail: vandev@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ"
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Книга 2 — Приложна математика и информатика
Том 89, 1995

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI"
FACULTE DE MATHEMATIQUES ET INFORMATIQUE
Livre 2 — Mathématiques Appliquée et Informatique
Tome 89, 1995

ON THE EFFECTIVE CONDUCTIVITY OF A CLASS OF RANDOM DISPERSIONS

KRASSIMIR D. ZVYATKOV

A new class of random dispersions is considered in which not only the location of the spheres is random, but their conductivity is random as well. The classical variational principles are employed in which classes of trial fields in the form of suitably truncated functional series are introduced. In this way three-point variational bounds on the effective conductivity of the dispersion are derived and discussed in more detail for some particular statistical distributions of sphere conductivity. A rigorous formula for the effective conductivity, correct to the order square of sphere fraction, is finally obtained which contains only absolutely convergent integrals.

Keywords: random media, effective properties, polydisperse structure.
1991 Mathematics Subject Classification: 73B35, 73S10.

1. INTRODUCTION

Consider a dispersion of *homogeneous* non-overlapping spheres of *random* conductivity $\widetilde{\kappa}_f$, immersed at random into an unbounded matrix of conductivity κ_m. For convenience of notations hereafter we represent the conductivity $\widetilde{\kappa}_f$ in the form $\widetilde{\kappa}_f = K_f \widetilde{s}$, where $K_f = \langle \widetilde{\kappa}_f \rangle$ is the mean conductivity of the sphere, embedded into the matrix. Then $\widetilde{s}$ represents their "non-dimensional conductivity" for which $\langle \widetilde{s} \rangle = 1$.

Let $\{\mathbf{x}_j\}$ be the random system of sphere's centers and at the position $\mathbf{x}_j$ a sphere with conductivity s_j, random as well, is centered. Thus a set of marked random points $\{\mathbf{x}_j, s_j\}$ is defined whose statistical description suffices for the dispersion. A similar marked random system was considered by Christov and Markov

[1, 2] in the study of dispersions of spheres with random radii $\widetilde{a}$. (For the general definition of sets of marked random points see [3].) We assume henceforth, for the sake of simplicity solely, that the spheres possess a fixed and non-random radius a. Then the random conductivity field $\kappa(\mathbf{x})$ of the dispersion has the form

$$\kappa(\mathbf{x}) = \kappa_m + \sum_j (K_f s_j - \kappa_m) h(\mathbf{x} - \mathbf{x}_j), \tag{1.1}$$

where $h(\mathbf{x})$ is the characteristic function for a single sphere located at the origin. In Sec. 2.1 we briefly discuss the statistical description of the system of marked random points $\{\mathbf{x}_j, s_j\}$, similar to that used in [1, 2].

For definiteness we shall deal with the problem of heat conduction through the random dispersion as a simple representative of a wide class of similar transport phenomena. The governing equations of the problem, in the absence of body sources, are

$$\nabla \cdot \mathbf{q}(\mathbf{x}) = 0, \quad \mathbf{q}(\mathbf{x}) = \kappa(\mathbf{x}) \nabla\theta(\mathbf{x}), \quad \langle \nabla\theta(\mathbf{x}) \rangle = \mathbf{G}, \tag{1.2}$$

where $\theta(\mathbf{x})$ is the random temperature field, $\mathbf{q}(\mathbf{x})$ — the heat flux vector, $\mathbf{G}$ is the prescribed macroscopic value of the temperature gradient, the brackets $\langle \cdot \rangle$ denote statistical averaging. Hereafter the media are assumed statistically homogeneous and isotropic. The solution of Eqs. (1.2) is understood in a statistical sense, so that one is to evaluate all multipoint moments (correlation functions) of $\theta(\mathbf{x})$ and the joint moments of $\kappa(\mathbf{x})$ and $\theta(\mathbf{x})$, see, e.g., [4]. Among the latter is the one-point moment

$$\langle \kappa(\mathbf{x}) \nabla\theta(\mathbf{x}) \rangle = \kappa^* \langle \nabla\theta(\mathbf{x}) \rangle = \kappa^* \mathbf{G}, \tag{1.3}$$

where κ^* is the effective conductivity of the medium.

As argued by Christov and Markov [5], the solution $\theta(\mathbf{x})$ of the random problem (1.2) can be expanded as a functional (Volterra-Wiener) series, generated by the conductivity field $\kappa(\mathbf{x})$, namely,

$$\theta(\mathbf{x}) = \mathbf{G} \cdot \mathbf{x} + \int K_1(\mathbf{x} - \mathbf{y}) \kappa'(\mathbf{y})\, d^3\mathbf{y}$$

$$+ \iint K_2(\mathbf{x} - \mathbf{y}_1, \mathbf{x} - \mathbf{y}_2) \left[\kappa'(\mathbf{y}_1)\kappa'(\mathbf{y}_2) - M_2^{\kappa}(\mathbf{y}_1 - \mathbf{y}_2)\right] d^3\mathbf{y}_1 d^3\mathbf{y}_2 + \cdots, \tag{1.4}$$

with certain non-random kernels T_i, $i = 1, 2, \ldots$ They also proposed to truncate this series afterwards. In Eq. (1.4) $\kappa'(\mathbf{x}) = \kappa(\mathbf{x}) - \langle \kappa \rangle$, $M_2^{\kappa}(\mathbf{x} - \mathbf{y}) = \langle \kappa'(\mathbf{x}) \kappa'(\mathbf{y}) \rangle$. (Hereafter the integrals with respect to spatial variables are over the whole $\mathbb{R}^3$ if the integration domain is not explicitly indicated.) Two types of applications for such truncated series could be envisaged. The first is to use them as approximate, in a certain sense, solutions of the problem (1.2). This possibility was discussed in more detail and worked out in the case of random dispersions of spheres by Markov [6, 7] and Markov and Christov [2]. For the dispersion under study this kind of application will be explained and worked out in Sec. 2.2. The second is to use such truncated series as classes of trial fields for the variational principles

[8, 12]. This idea was developed by Markov [8] on the base of the classical principle, corresponding to the problem (1.2), namely,

$$W_A[\theta(\cdot)] = \left\langle \kappa(\mathbf{x})|\nabla\theta(\mathbf{x})|^2 \right\rangle \longrightarrow \min, \quad \langle \nabla\theta(\mathbf{x}) \rangle = \mathbf{G}, \tag{1.5}$$

$\min W_A = \kappa^* G^2$, see, e.g., [4]. For example, the simplest non-trivial class is obtained when the functional series (1.4) is truncated after the single integral term, i.e.

$$\mathcal{K}_A^{(1)} = \left\{ \theta(\mathbf{x}) \mid \theta(\mathbf{x}) = \mathbf{G}\cdot\mathbf{x} + \int K_1(\mathbf{x}-\mathbf{y})\kappa'(\mathbf{y})\, d^3\mathbf{y} \right\}, \tag{1.6}$$

where $K_1(\mathbf{x})$ is an adjustable kernel. This class was introduced and discussed in detail by Markov [8], where it was shown that minimizing $W_A[\theta(\cdot)]$ over the class $\mathcal{K}_A^{(1)}$ gives the best three-point upper bound $\kappa^{(3)}$ on the effective conductivity κ^*, i.e. the most restrictive one which uses three-point statistical information for the medium. In order to obtain the appropriate three-point lower bound on κ^*, it is necessary to consider the classical dual variational principle for the problem (1.2) formulated with respect to the heat flux $\mathbf{q}(\mathbf{x}) = \nabla \times \Phi(\mathbf{x})$,

$$W_B[\Phi(\cdot)] = \left\langle k(\mathbf{x})|\nabla \times \Phi(\mathbf{x})|^2 \right\rangle \longrightarrow \min, \quad \langle \mathbf{q}(\mathbf{x}) \rangle = \mathbf{Q}, \tag{1.7}$$

$\min W_B = k^* Q^2$ (here $k(\mathbf{x}) = 1/\kappa(\mathbf{x})$ and $k^* = 1/\kappa^*$), over a class of the kind (1.6). In Sec. 3.1 we shall derive the optimal three-point bounds for the dispersion making use of an alternative variational procedure successfully applied in the monodisperse case, see [8, 9, 12].

Moreover, Markov [8] showed how the earlier proposed variational techniques could be put into this general frame. For example, the Beran method [13] is a Ritz type procedure in which the kernel K_1 in (1.6) is chosen to be proportional to the fixed (Beran's) kernel K_B:

$$K_1(\mathbf{x}) = \lambda K_B(\mathbf{x}), \quad K_B(\mathbf{x}) = \mathbf{G}\cdot\nabla\frac{1}{4\pi|\mathbf{x}|}, \tag{1.8}$$

where $\lambda \in \mathbb{R}$ is an adjustable parameter. The question of the optimality of Beran's procedure for the dispersion under study will be discussed in Sec. 3.2. It will be shown that it is not optimal even to the order c, where c is the volume fraction of the spheres. Finally, in Sec. 4, using some of the author's ideas of his recent work [14], an exact c^2-formula for the effective conductivity κ^* of the dispersion under study will be found in a variational way.

2. STATISTICAL DESCRIPTION OF THE DISPERSION AND FACTORIAL FUNCTIONAL EXPANSION

2.1. STATISTICAL DESCRIPTION OF THE DISPERSION

The system of marked random points $\{\mathbf{x}_j, s_j\}$ can be considered as a set of points randomly distributed in the four-dimensional domain $\mathbb{R}^3 \times \mathbb{U}$, where $\mathbb{U} = (0, +\infty)$. Similarly to the monodisperse case, this system is fully described

by the multipoint probability densities $F_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; s_1, \ldots, s_n)$, see, e.g., [1, 2, 8]. The latter define the probability

$$dP = F_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; s_1, \ldots, s_n)\, d^3\mathbf{y}_1 \ldots d^3\mathbf{y}_n\, ds_1 \ldots ds_n \tag{2.1}$$

to find simultaneously a center of sphere within the infinitesimal volumes

$$\mathbf{y}_i \le \mathbf{y} < \mathbf{y}_i + d\mathbf{y}_i \tag{2.2a}$$

of the spatial positions $\mathbf{y}_i$ with conductivities $\widetilde{s}_1, \ldots, \widetilde{s}_n$ in the vicinities

$$s_i \le \widetilde{s} < s_i + ds_i \tag{2.2b}$$

of the values $s_1, \ldots, s_n$, respectively, $i = 1, \ldots, n$.

The functions $F_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; s_1, \ldots, s_n)$ define too rich a class of dispersions whose study seems very complicated in general. That is why, if our aim is to reach certain tangible results, one must narrow this class. The following arguments lead in a natural way to such a simplification. Let dP_Y be the probability to find simultaneously a center of sphere in each of the volumes (2.2a), regardless to the conductivity of the latter. Obviously, $dP \le dP_Y$ and

$$dP_Y = f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)\, d^3\mathbf{y}_1 \ldots d^3\mathbf{y}_n, \tag{2.3}$$

where the functions $f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ are the multipoint probability densities for the system of non-marked random points $\mathbf{x}_j$, i.e. they are the same that appear in the monodisperse case, see, e.g., [8]. Then $dP = dP_Y dP^*$, where dP^* is the conditional probability, namely, the probability to find simultaneously a center of sphere in the volumes (2.2a) with conductivities $\widetilde{s}_1, \ldots, \widetilde{s}_n$ in the regions (2.2b) respectively, provided a center of sphere is found in the volumes (2.2a). Hence

$$dP^* = \eta_n(s_1, \ldots, s_n \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)\, ds_1 \ldots ds_n,$$

where

$$F_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; s_1, \ldots, s_n) = f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)\, \eta_n(s_1, \ldots, s_n \mid \mathbf{y}_1, \ldots, \mathbf{y}_n), \tag{2.4}$$

$n = 1, 2 \ldots$ Obviously, the dependence of functions η_n upon $\mathbf{y}_1, \ldots, \mathbf{y}_n$ reflects the "selectivity" of these sphere's locations toward spheres of certain conductivities. The consideration of dispersions in the general case, when such a "selectivity" is arbitrary, seems a hopeless problem. That is why we adopt now the following simplifying assumption concerning the structure of the dispersions: *There exist no locations in the space* $\mathbb{R}^3$ *which possess selectivity toward spheres of certain conductivities.* Hence we assume that

$$\eta_n(s_1, \ldots, s_n \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) = P_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)$$

or, according to (2.4),

$$F_n(\mathbf{y}_1, \ldots, \mathbf{y}_n; s_1, \ldots, s_n) = f_n(\mathbf{y}_1, \ldots, \mathbf{y}_n)\, P_n(s_1, \ldots, s_n), \tag{2.5}$$

which means, as a matter of fact, that there is no correlation between location and conductivity of the spheres. The functions $P_n(s_1, \ldots, s_n)$ are the multivariate probability densities of conductivities of spheres, regardless to the spatial positions of the latter; they give the probability dP_S^* of n arbitrarily chosen spheres of the dispersion, having conductivities in the vicinities (2.2b), to be $dP_S^* = P_n(s_1, \ldots, s_n)\, ds_1 \ldots ds_n$.

Since the dispersions under study are assumed statistically homogeneous and isotropic, the system $\{\mathbf{x}_j\}$ has the same properties. Hence, in particular, $f_1 = n$ and $f_k = f_k(\mathbf{y}_{2,1}, \ldots, \mathbf{y}_{k,1})$, where $\mathbf{y}_{i,j} = \mathbf{y}_j - \mathbf{y}_i$ and n denotes the number density, i.e. the mean number of points $\mathbf{x}_j$ per unit volume. Obviously, $n = c/V_a$, where $V_a = \frac{4}{3}\pi a^3$ is the volume of a single sphere. Moreover, we shall assume, as usual, that $f_k \sim n^k$, i.e. f_k has the asymptotic order n^k at $n \to 0$, $k = 1, 2, \ldots$, see [8]. We shall note also that the assumption of non-overlapping of spheres yields

$$f_k(\mathbf{y}_1, \ldots, \mathbf{y}_k) = 0, \quad \text{if} \quad |\mathbf{y}_i - \mathbf{y}_j| < 2a \quad \text{for a pair} \quad i \neq j\,.$$

Taking into account this assumption and (2.5) for the first pair of probability densities F_1 and F_2 we have

$$F_1(\mathbf{y}; s) = nP(s), \quad F_2(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2) = n^2 g_0(r) P_2(s_1, s_2), \tag{2.6}$$

where $P(s) = P_1(s)$, $r = |\mathbf{y}_2 - \mathbf{y}_1|$ and g_0 is the zero-density limit of the well-known radial distribution function $g(r) = f_2(r)/n^2$, i.e. $g(r) = g_0(r) + O(n)$.

Let

$$\psi(\mathbf{x}; s) = \sum_j \delta(\mathbf{x} - \mathbf{x}_j)\delta(s - s_j) \tag{2.7}$$

be the Stratonovich random density field generated by the system of marked random points $\{\mathbf{x}_j, s_j\}$ (see [15, 2]). According to Eq. (1.1) the field $\kappa(\mathbf{x})$ can be written then as

$$\kappa(\mathbf{x}) = \langle\kappa\rangle + \int\!\!\int (K_f s - \kappa_m) h(\mathbf{x} - \mathbf{y}) \psi'(\mathbf{y}; s)\, d^3\mathbf{y}\, ds, \tag{2.8}$$

where $\psi'(\mathbf{y}; s)$ is the fluctuating part of the field $\psi(\mathbf{y}; s)$. (Hereafter the integrals with respect to the mark s are over the semiaxis $(0, +\infty)$.) The random field $\psi(\mathbf{x}; s)$ is uniquely defined by the random set $\{\mathbf{x}_j, s_j\}$ and vice versa. In particular, the multipoint moments of $\psi(\mathbf{x}; s)$ can easily be expressed by means of the probability densities F_k:

$$\begin{gathered}
\langle\psi(\mathbf{y}; s)\rangle = F_1(\mathbf{y}; s) = nP(s), \\
\langle\psi(\mathbf{y}_1; s_1)\psi(\mathbf{y}_2; s_2)\rangle = F_1(\mathbf{y}_1; s_1)\delta(\mathbf{y}_{1,2})\delta(s_{1,2}) + F_2(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2), \\
\langle\psi(\mathbf{y}_1; s_1)\psi(\mathbf{y}_2; s_2)\psi(\mathbf{y}_3; s_3)\rangle = F_1(\mathbf{y}_1; s_1)\delta(\mathbf{y}_{1,2})\delta(s_{1,2})\delta(\mathbf{y}_{1,3})\delta(s_{1,3}) \\
+3\{\delta(\mathbf{y}_{1,3})\delta(s_{1,3})F_2(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2)\}_s + F_3(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3; s_1, s_2, s_3),
\end{gathered} \tag{2.9}$$

etc., see [1, 2], where $\{\cdot\}_s$ denotes symmetrization with respect to all different combinations of indices in the brackets, $s_{i,j} = s_j - s_i$.

Similarly to the considerations in [6, 7] (for monodisperse case) and [2] (for the dispersion of spheres with random radii), it is reasonable to develop the random temperature field $\theta(\mathbf{x})$ in the following functional series

$$\theta(\mathbf{x}) = T_0(\mathbf{x}) + \int\int T_1(\mathbf{x}-\mathbf{y}, s)\Delta_\psi^{(1)}(\mathbf{y}; s)\, d^3\mathbf{y} ds$$

$$+\int\int\int\int T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2)\Delta^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2)\, d^3\mathbf{x}_1\, d^3\mathbf{y}_2\, ds_1 ds_2 + \cdots, \quad (2.10)$$

where

$$\Delta_\psi^{(0)} = 1, \quad \Delta_\psi^{(1)}(\mathbf{y}; s) = \psi(\mathbf{y}; s),$$

$$\Delta_\psi^{(k)}(\mathbf{y}_1, \ldots, \mathbf{y}_k; s_1, \ldots, s_k) = \psi(\mathbf{y}_1; s_1)[\psi(\mathbf{y}_2; s_2) - \delta(\mathbf{y}_{2,1})\delta(s_{2,1})] \quad (2.11)$$

$$\ldots[\psi(\mathbf{y}_k; s_k) - \delta(\mathbf{y}_{k,1})\delta(s_{k,1}) - \cdots - \delta(\mathbf{y}_{k,k-1})\delta(s_{k,k-1})], \quad k = 2, 3, \ldots,$$

are the random fields, generated by the random density field $\psi(\mathbf{x}; s)$, and called in [7] factorial fields. The kernel T_k in (2.10) can be easily expressed by means of the first k kernels of the series (1.3). According to a basic result of [7], the series (2.10) is virial in the sense that the truncation after the p-tuple term of its gives results for all multipoint moments of the solution $\theta(\mathbf{x})$ to the random problem (1.2), which are correct to the order c^p provided the first kernels T_i, $i = 0, \ldots, p$, are properly identified. A general procedure for the identification of the kernels T_i is described in [2, 6, 7].

Since our aim is the evaluation of the effective conductivity κ^* to the order c^2, we are interested in the solution of the problem (1.2) to the same order. To simplify the analysis, after [2, 6, 7] we render the series (2.10) n^2-orthogonal in the sense that the averaged value of the product of any pair of its different terms has the order $o(n^2)$. To this end we introduce the following linear combinations of the factorial fields (2.11):

$$D_\psi^{(0)} = 1, \quad D_\psi^{(1)}(\mathbf{y}; s) = \Delta_\psi^{(1)}(\mathbf{y}; s) - nP(s) = \psi'(\mathbf{y}; s),$$

$$D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2) = \Delta_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2) - n^2 g_0(\mathbf{y}_{2,1})P_2(s_1, s_2)$$

$$-n^2 g_0(\mathbf{y}_{2,1})P_2(s_1, s_2)[D_\psi^{(1)}(\mathbf{y}_1; s_1)/P(s_1) + D_\psi^{(1)}(\mathbf{y}_2; s_2)/P(s_2)],$$

$$D_\psi^{(k)}(\mathbf{y}_1, \ldots, \mathbf{y}_j; s_1, \ldots, s_k) = \Delta_\psi^{(k)}(\mathbf{y}_1, \ldots \mathbf{y}_k; s_1, \ldots, s_k), \quad (2.12)$$

$k = 3, 4, \ldots$ As a consequence of Eqs. (2.9) and (2.11) it can be easily verified that

$$\left\langle D_\psi^{(1)}(\mathbf{y}; s)\right\rangle = 0, \quad \left\langle D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2)\right\rangle = o(n^2), \quad (2.13a)$$

$$\left\langle D_\psi^{(1)}(\mathbf{y}_1; s_1) D_\psi^{(2)}(\mathbf{y}_2, \mathbf{y}_3; s_2, s_3)\right\rangle = o(n^2). \quad (2.13b)$$

Since the series (2.10) is virial, these relations suffice to claim that the fields (2.12) form an n^2-orthogonal system. Then let us truncate the series (2.10) after the four-tuple integral term. Thus we obtain the kind of the c^2-solution of the random problem (1.2) for the dispersion. In the truncated series we rearrange the terms in such a manner that only the n^2-orthogonal fields $D_\psi^{(1)}$ and $D_\psi^{(2)}$ enter:

$$\theta(\mathbf{x}) = \mathbf{G}\cdot\mathbf{x} + \int\!\!\int T_1(\mathbf{x}-\mathbf{y}, s) D_\psi^{(1)}(\mathbf{y}; s)\, d^3\mathbf{y} ds$$

$$+ \int\!\!\int\!\!\int\!\!\int T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2) D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2)\, d^3\mathbf{y}_1 d^3\mathbf{y}_2\, ds_1 ds_2. \quad (2.14)$$

The new kernels T_1 and T_2 here (no new notations are used for them) are linear combinations of the kernels T_0, T_1 and T_2 of the series (2.10). The zeroth-order term in (2.14) is indeed $\mathbf{G}\cdot\mathbf{x}$, since $D_\psi^{(1)}$ and $D_\psi^{(2)}$ are centered and $\langle\nabla\theta(\mathbf{x})\rangle = \mathbf{G}$, see Eqs. (2.13) and (1.2).

The identification of the kernels T_1 and T_2 can be performed by a procedure, proposed originally by Christov and Markov [5], see also [2, 6, 7]. It consists in inserting the truncated series (2.14) into the random equation (1.2), multiplying the result by the fields $D_\psi^{(p)}$, $p = 0, 1, 2$, and averaging the results. In this way a certain system of integral-differential equations for the needed kernels of the truncated series can be straightforwardly derived. Here we employ an alternative method, recently proposed in [14] for the monodisperse case. Namely, the truncated series (2.14) will be inserted into the classical variational principle (1.5) as a class of trial fields, varying the kernels. Since this class contains the actual temperature field to the order c^2, the obtained equations for the optimal kernels T_1 and T_2 are the same as those for the needed kernels in (2.14). In particular, this procedure leads in passing to the exact determination of the effective conductivity to the order c^2.

In what follows we shall need also the following formulae for the moments of the field (2.12):

$$\left\langle D_\psi^{(1)}(\mathbf{y}_1; s_1) D_\psi^{(1)}(\mathbf{y}_2; s_2)\right\rangle = nP(s_1)\delta(\mathbf{y}_{1,2})\delta(s_{1,2}) - n^2\mathcal{R}_0(\mathbf{y}_{1,2}; s_1, s_2), \quad (2.15a)$$

$$\left\langle D_\psi^{(1)}(\mathbf{y}_1; s_1) D_\psi^{(1)}(\mathbf{y}_2; s_2) D_\psi^{(1)}(\mathbf{y}_3; s_3)\right\rangle = nP(s_1)\delta(\mathbf{y}_{1,2})\delta(s_{1,2})\delta(\mathbf{y}_{1,3})\delta(s_{1,3})$$
$$-n^2 3\{\delta(\mathbf{y}_{1,2})\delta(s_{1,2})\mathcal{R}_0(\mathbf{y}_{2,3}; s_2, s_3)\}_s, \quad (2.15b)$$

$$\left\langle D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2) D_\psi^{(1)}(\mathbf{y}_3; s_3) D_\psi^{(1)}(\mathbf{y}_4; s_4)\right\rangle$$
$$= \left\langle D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2) D_\psi^{(2)}(\mathbf{y}_3, \mathbf{y}_4; s_3, s_4)\right\rangle = n^2 g_0(\mathbf{y}_{2,1}) P_2(s_1, s_2) \quad (2.15c)$$
$$\times [\delta(\mathbf{y}_{3,1})\delta(s_{3,1})\delta(\mathbf{y}_{4,2})\delta(s_{4,2}) + \delta(\mathbf{y}_{4,1})\delta(s_{4,1})\delta(\mathbf{y}_{3,2})\delta(s_{3,2})],$$

$$\left\langle D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2) D_\psi^{(2)}(\mathbf{y}_3, \mathbf{y}_4; s_3, s_4) D_\psi^{(1)}(\mathbf{y}_5; s_5)\right\rangle$$
$$= n^2 g_0(\mathbf{y}_{2,1}) P_2(s_1, s_2)[\delta(\mathbf{y}_{5,1})\delta(s_{5,1}) + \delta(\mathbf{y}_{5,2})\delta(s_{5,2})] \quad (2.15d)$$
$$\times [\delta(\mathbf{y}_{3,1})\delta(s_{3,1})\delta(\mathbf{y}_{4,2})\delta(s_{4,2}) + \delta(\mathbf{y}_{4,1})\delta(s_{4,1})\delta(\mathbf{y}_{3,2})\delta(s_{3,2})],$$

where

$$\mathcal{R}_0(\mathbf{y}_{2,1}; s_1, s_2) = P(s_1)P(s_2) - g_0(|\mathbf{y}_{2,1}|)P_2(s_1, s_2)\,; \tag{2.16}$$

they are correct to the order n^2 and represent straightforward consequences of Eqs. (2.9), (2.11) and (2.12).

3. VARIATIONAL THREE-POINT BOUNDS

3.1. THE OPTIMAL THREE-POINT BOUNDS FOR THE DISPERSION

It is natural to begin the consideration of the classical variational principle (1.5) on the simpler class of trial fields that it yields when the factorial series (2.14) is truncated after the one-tuple integral term. Namely, we introduce the class

$$T_A^{(1)} = \left\{ \theta(\mathbf{x}) \;\middle|\; \theta(\mathbf{x}) = \mathbf{G}\cdot\mathbf{x} + \int\!\!\int T_1(\mathbf{x}-\mathbf{y}, s) D_\psi^{(1)}(\mathbf{y}; s)\, d^3\mathbf{y}\, ds \right\}, \tag{3.1}$$

where $T_1(\mathbf{x}, s)$ is an adjustable kernel. Obviously, this class contains the actual temperature field to the order c only. That is why one can obtain the exact value of effective conductivity κ^* to the same order only, together with certain bounds on κ^* for the higher order of c.

This class is the counterpart of the class (1.6). Due to Eq. (2.8), the classes (1.6) and (3.1) coincide: if a transition from $\kappa'(\mathbf{x})$ to $\psi'(\mathbf{x})$ is performed according to Eq. (2.8), the kernel $K_1(\mathbf{x})$ is transformed into the kernel $T_1(\mathbf{x}, s)$ by means of the convolution with the characteristic function $h(\mathbf{x})$:

$$T_1(\mathbf{x}, s) = (K_f s - \kappa_m)\,(h * K_1)\,(\mathbf{x}) = (K_f s - \kappa_m) \int h(\mathbf{x}-\mathbf{y}) K_1(\mathbf{y})\, d^3\mathbf{y}. \tag{3.2}$$

Consequently, the upper bound on κ^*, obtained from the restriction of the functional W_A over the class $T_A^{(1)}$ coincides with the optimal third-order bound $\kappa^{(3)}$, see Sec. 1. Moreover, due to Eqs. (2.8) and (2.9), we can claim that the bound $\kappa^{(3)}$ is the best one for the dispersion which employs the statistical information provided by the two- and three-point probability densities F_2 and F_3.

Making use of Eq. (2.8) and the formulae (2.15) for the moments of the fields $D_\psi^{(1)}$, the restriction $W_A^{(1)}[T_1(\cdot)]$ of the functional W_A over the class (3.1) becomes

$$W_A^{(1)}[T_1(\cdot)] = W_A\,\Big|_{T_A^{(1)}} = \langle\kappa\rangle\, \mathbf{G}^2 + n\,\langle\kappa\rangle \left\{ \int\!\!\int |\nabla T_1(\mathbf{z}, s)|^2 P(s)\, d^3\mathbf{z}\, ds \right.$$

$$\left. - n \int\!\!\int\!\!\int\!\!\int \mathcal{R}_0(\mathbf{z}_1 - \mathbf{z}_2; s_1, s_2) \nabla T_1(\mathbf{z}_1, s_1)\cdot\nabla T_1(\mathbf{z}_2, s_2)\, d^3\mathbf{z}_1\, d^3\mathbf{z}_2\, ds_1\, ds_2 \right\}$$

$$+ 2n\,\mathbf{G}\cdot\left\{ \int\!\!\int (K_f s - \kappa_m) h(\mathbf{z}) \nabla T_1(\mathbf{z}, s) P(s)\, d^3\mathbf{z}\, ds \right.$$

$$\left. - n \int\!\!\int\!\!\int (K_f s_1 - \kappa_m) \mathcal{F}_0(\mathbf{z}; s_1, s_2) \nabla T_1(\mathbf{z}, s_2)\, d^3\mathbf{z}\, ds_1\, ds_2 \right\}$$

$$+n\left\{\int(K_f s-\kappa_m)h(\mathbf{z})|\nabla T_1(\mathbf{z},s)|^2P(s)\,d^3\mathbf{z}\,ds\right.$$

$$-n\left[2\iiiint(K_f s_1-\kappa_m)h(\mathbf{z}_1)\mathcal{R}_0(\mathbf{z}_{2,1},s_1,s_2)\nabla T_1(\mathbf{z}_1,s_1)\right.$$

$$\cdot\nabla T_1(\mathbf{z}_2,s_2)\,d^3\mathbf{z}_1 d^3\mathbf{z}_2\,ds_1 ds_2$$

$$\left.\left.+\iiint\mathcal{F}_0(\mathbf{z};s_1,s_2)|\nabla T_1(\mathbf{z},s_2)|^2\,d^3\mathbf{z}\,ds_1\,ds_2\right]\right\}+o(n^2), \tag{3.3}$$

where

$$\mathcal{F}_0(\mathbf{z};s_1,s_2)=\int h(\mathbf{y})\mathcal{R}_0(\mathbf{z}-\mathbf{y};s_1,s_2)\,d\mathbf{y}. \tag{3.4}$$

Hereafter the differentiation is with respect to the appropriate spatial variable.

The optimal kernel $T_1(\mathbf{x},s)$, i.e. the solution of the Euler-Lagrange equation for the functional $W_A^{(1)}$, is looked for in the virial form

$$T_1(\mathbf{x},s)=T_1(\mathbf{x},s;n)=T_{1,0}(\mathbf{x},s)+T_{1,1}(\mathbf{x},s)\,n+\cdots \tag{3.5}$$

This representation of $T_1(\mathbf{x},s)$ induces the appropriate virial expansion of the functional (3.3):

$$W_A^{(1)}[T_1(\cdot)]=\langle\kappa\rangle G^2+W_A^{(1,1)}[T_{1,0}(\cdot)]n+W_A^{(1,2)}[T_{1,0}(\cdot),T_{1,1}(\cdot)]n^2+\cdots \tag{3.6}$$

The functionals $W_A^{(1,1)}$ and $W_A^{(1,2)}$ depend on the indicated virial coefficients as follows:

$$W_A^{(1,1)}[T_{1,0}(\cdot)]=\kappa_m\iint|\nabla T_{1,0}(\mathbf{x},s)|^2P(s)\,d^3\mathbf{x}\,ds$$

$$+\iint(K_f s-\kappa_m)h(\mathbf{x})[\,\nabla T_{1,0}(\mathbf{x},s)+2\mathbf{G}\,]\cdot\nabla T_{1,0}(\mathbf{x},s)P(s)\,d^3\mathbf{x}\,ds, \tag{3.7}$$

$$W_A^{(1,2)}[T_{1,0}(\cdot),T_{1,1}(\cdot)]=\overline{W}_A^{(1,2)}[T_{1,0}(\cdot)]+2\int P(s)\,ds$$

$$\times\int\nabla\cdot\{\kappa_m\nabla T_{1,0}(\mathbf{x},s)+(K_f s-\kappa_m)h(\mathbf{x})[\,\mathbf{G}+\nabla T_{1,0}(\mathbf{x},s)\,]\}\,T_{1,1}(\mathbf{x},s)\,d^3\mathbf{x}, \tag{3.8a}$$

$$\overline{W}_A^{(1,2)}[T_{1,0}(\cdot)]=(K_f-\kappa_m)V_a\iint|\nabla T_{1,0}(\mathbf{x},s)|^2P(s)\,d^3\mathbf{x}\,ds$$

$$-\iiiint(K_f s_1-\kappa_m)h(\mathbf{x}_1)|\nabla T_{1,0}(\mathbf{x}_2,s_2)|^2\mathcal{R}_0(\mathbf{x}_1-\mathbf{x}_2;s_1,s_2)\,d^3\mathbf{x}_1\,d^3\mathbf{x}_2\,ds_1\,ds_2$$

$$+\kappa_m\iiiint\nabla T_{1,0}(\mathbf{x}_1,s_1)\cdot\nabla T_{1,0}(\mathbf{x}_2,s_2)\mathcal{R}_0(\mathbf{x}_1-\mathbf{x}_2;s_1,s_2)\,d^3\mathbf{x}_1\,d^3\mathbf{x}_2\,ds_1\,ds_2$$

$$-2\iint\{\kappa_m\nabla T_{1,0}(\mathbf{x}_1,s_1)+(K_f s_1-\kappa_m)h(\mathbf{x}_1)[\,\mathbf{G}+\nabla T_{1,0}(\mathbf{x}_1,s_1)\,]\}\,d^3\mathbf{x}_1\,ds_1$$

$$\cdot \iint \nabla T_{1,0}(\mathbf{x}_2, s_2) \mathcal{R}_0(\mathbf{x}_1 - \mathbf{x}_2; s_1, s_2)\, d^3\mathbf{x}_2\, ds_2. \tag{3.8b}$$

The optimal kernel $T_1(\mathbf{x}, s)$ satisfies the equation $\delta W_A^{(1)} = 0$, so that we have, in particular,

$$\delta W_A^{(1,1)}[T_{1,0}(\cdot)] = 0, \quad \delta W_A^{(1,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot)] = 0. \tag{3.9}$$

The first of these equations yields straightforwardly

$$P(s)\nabla \cdot \{\kappa_m \nabla T_{1,0}(\mathbf{x}, s) + (K_f s - \kappa_m) h(\mathbf{x})[\,\mathbf{G} + \nabla T_{1,0}(\mathbf{x}, s)\,]\} = 0, \tag{3.10}$$

which is just the equation for the disturbance, $T^{(1)}(\mathbf{x}, s)$, to the temperature field $\mathbf{G} \cdot \mathbf{x}$ in an unbounded matrix, introduced by a single spherical inhomogeneity of conductivity $K_f s$, located at the origin. The analytic form of this disturbance is well-known:

$$T_{1,0}(\mathbf{x}, s) = T^{(1)}(\mathbf{x}, s) = 3\beta(s)\, \mathbf{G} \cdot \nabla \varphi(\mathbf{x}), \quad \beta(s) = \frac{K_f s - \kappa_m}{K_f s + 2\kappa_m}; \tag{3.11}$$

here

$$\varphi(\mathbf{x}) = h * \frac{1}{4\pi|\mathbf{x}|}, \quad \text{i.e.} \quad \varphi(\mathbf{x}) = \int \frac{h(\mathbf{y})}{4\pi|\mathbf{x} - \mathbf{y}|}\, d^3\mathbf{y}$$

is the Newtonian potential for a single sphere of radius a, located at the origin. (We assume, obviously enough, that $P(s) \neq 0$.)

With $T_{1,0}(\mathbf{x}, s)$ already found, one should vary only $T_{1,1}(\mathbf{x}, s)$ in the functional (3.8) in order to derive the Euler-Lagrange equation for the latter. However, this is not possible, because Eq. (3.10) yields

$$W_A^{(1,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot)] = \overline{W}_A^{(1,2)}[T_{1,0}(\cdot)]. \tag{3.12}$$

Hence, according to Eq. (3.6) for the optimal upper three-point bound $\kappa^{(3)}$ we have

$$\kappa^* G^2 \leq \kappa^{(3)} G^2 = \langle \kappa \rangle\, G^2 + \frac{1}{V_a} W_A^{(1,1)}[T_{1,0}(\cdot)]c + \frac{1}{V_a} \overline{W}_A^{(1,2)}[T_{1,0}(\cdot)]c^2 + o(c^2). \tag{3.13}$$

The foregoing reasoning has two implications. First, we can conclude that the optimal upper bound $\kappa^{(3)}$ to the order c^2 depends only on the field $T^{(1)}(\mathbf{x}, s)$; the explicit form of $T_{1,1}(\mathbf{x}, s)$ is not required at all, see Eq. (3.13). Second, the kernel $T_1(\mathbf{x}, s)$ is optimal to the order c^2 if its leading coefficient $T_{1,0}(\mathbf{x}, s)$ in the virial expansion (3.5) is proportional to the single-sphere disturbance field $T^{(1)}(\mathbf{x}, s)$. In this connection it is to be noted that the known Ritz type procedure of Torquato [16] leads to the optimal bound to the order c^2. (For the latter the kernel $T_1(\mathbf{x}, s)$ in (3.1) should be chosen as $T_1(\mathbf{x}, s) = \lambda T^{(1)}(\mathbf{x}, s)$, where λ is an adjustable parameter.) This fact holds also for dispersions of radial inhomogeneous spheres with random radii, see [17]. To the order c^p at $p > 2$, however, the cluster bounds of Torquato are not optimal even for the monodisperse case, see [9, 12].

Repeating the above arguments with respect to the dual principle (1.7) leads to a fully similar conclusion for the optimal lower bound, namely, that to the order c^2 the latter is fully determined by the disturbance $\mathbf{q}^{(1)}(\mathbf{x}, s)$ to the heat flux $\mathbf{Q}$ in an unbounded matrix, introduced by a single spherical inhomogeneity of conductivity $K_f s$, located at the origin.

Let

$$\frac{\kappa^*}{\kappa_m} = 1 + a_{1\kappa}c + a_{2\kappa}c^2 + \cdots \tag{3.14}$$

be the virial expansion for the effective conductivity of the dispersion. Making use of Eqs. (3.10) and (3.11), the connection of the disturbances $T^{(1)}(\mathbf{x}, s)$ and $\mathbf{q}^{(1)}(\mathbf{x}, s)$, and the relation $h(\mathbf{x})\nabla T^{(1)}(\mathbf{x}, s) = -\beta(s)h(\mathbf{x})\mathbf{G}$, we easily get as a consequence of (3.7), (3.13) and their counterparts for the dual variational principle (1.7), that

$$a_{1\kappa} = 3N, \quad N = N[P(\cdot)] = \langle \beta(\widetilde{s}) \rangle = \int_0^\infty \beta(s)P(s)\,ds, \tag{3.15}$$

so that the upper and lower bounds coincide to the order c, as it should have been expected. After simple algebra, based on Eqs. (3.8), (3.10), (3.11), (3.13) and their counterparts for the lower bound, we get the following inequalities for the c^2-coefficient $a_{2\kappa}$:

$$a^l_{2\kappa} \le a_{2\kappa} \le a^u_{2\kappa}, \tag{3.16a}$$

$$a^l_{2\kappa} = 3\left\{N^2 + \int_0^\infty \frac{K_f s_1 - \kappa_m}{K_f s_1} ds_1 \int_0^\infty \beta^2(s_2)\mathcal{M}_2(s_1, s_2)ds_2\right\}, \tag{3.16b}$$

$$a^u_{2\kappa} = 3\left\{N^2 + \int_0^\infty \frac{K_f s_1 - \kappa_m}{\kappa_m} ds_1 \int_0^\infty \beta^2(s_2)\mathcal{M}_2(s_1, s_2)ds_2\right\}, \tag{3.16c}$$

where

$$\mathcal{M}_2(s_1, s_2) = -\frac{1}{2\pi}\int_0^\infty \frac{1}{r^3}\frac{\partial}{\partial r}\mathcal{F}_0(r; s_1, s_2)\,dr \tag{3.17}$$

is a statistical parameter for the dispersion; the function $\mathcal{F}_0(r; s_1, s_2)$ is defined in Eq. (3.4), $r = |x|$.

The formula (3.15) clearly indicates that the effective conductivity κ^* depends on the statistical distribution of conductivity of spheres even to the order c. (Let us recall that the c-coefficient $a_{1\kappa}$ is independent of the size distribution for a dispersion of spheres of random radii, see [1, 17, 18].) Moreover, it is to be noted that in general $\langle \beta(\widetilde{s}) \rangle \neq \beta(\langle \widetilde{s} \rangle) = \beta(1) = (K_f - \kappa_m)/(K_f + 2\kappa_m)$, so that the dispersion is *not equivalent* to a monodisperse dispersion of sphere (with the same sphere fraction c) of the mean conductivity K_f even to the order c, see Eq. (3.15).

According to (3.2), the Beran's kernel $K_B(\mathbf{x})$, see (1.8), is transformed into the kernel

$$T_B(\mathbf{x}, s) = (K_f s - \kappa_m)\, \mathbf{G} \cdot \nabla\varphi(\mathbf{x}) \tag{3.18}$$

at the transition from $\kappa'(\mathbf{x})$ to $\psi'(\mathbf{x}; s)$. Due to Eq. (3.11), the kernel $K_1(\mathbf{x}) = \lambda K_B(\mathbf{x})$, $\lambda \in \mathbb{R}$, will be optimal to the order c and consequently to the order c^2 also (see (3.12)), if and only if $T^{(1)}(\mathbf{x}, s) = \lambda T_B(\mathbf{x}, s)$ for a certain $\lambda \in \mathbb{R}$ and *for all* s such that $P(s) \neq 0$. It is shown, however, that it is possible only if $P(s) = \delta(s - s_0)$, i.e., if the probability to find a sphere of conductivity different of $K_f s_0$ is equal to zero; in other words, for the usually considered dispersions of spheres possessing one and the same conductivity. Hence, we can conclude that the Beran's bounds are not optimal even to the order c for the considered dispersions.

The above arguments imply the following simple way for a generalization of the Beran's procedure. Namely, if we choose the kernel $K_1(\mathbf{x})$ in the form $K_1(\mathbf{x}) = \lambda(s) K_B(\mathbf{x})$, see (1.8), where now $\lambda(s)$ is an adjustable function, then the optimal bounds to the order c^2 will be obtained.

Let us note that the Beran's bounds are more complicated for the dispersion under study. For example, the minimization of the functional (3.7) at $T_{1,0}(\mathbf{x}, s) = \lambda T_B(\mathbf{x}, s)$ with respect to $\lambda \in \mathbb{R}$ leads to the following upper bound $a^u_{1\kappa(B)}$ on the c-coefficient $a_{1\kappa}$:

$$a_{1\kappa} \leq a^u_{1\kappa(B)}\,, \quad a^u_{1\kappa(B)} = \frac{K_f}{\kappa_m} - 1 - \frac{\left\langle (K_f \widetilde{s} - \kappa_m)^2 \right\rangle^2}{\kappa_m \left\langle (K_f \widetilde{s} - \kappa_m)^2 (K_f \widetilde{s} + 2\kappa_m) \right\rangle}\,; \tag{3.19}$$

the equality sign, $a_{1\kappa} = a^u_{1\kappa(B)}$, being achieved at $P(s) = \delta(s - s_0)$ only, i.e. if the spheres have one and the same conductivity.

3.3. EXAMPLES

We shall illustrate the influence of the statistical distribution of conductivities of spheres on the obtained c^2-bounds (3.15), (3.16). First, we note that if we adopt the assumption of statistical independence of the conductivities of each two spheres, i.e.

$$P_2(s_1, s_2) = P(s_1) P(s_2), \tag{3.20}$$

which sounds reasonable enough (at least in the dilute case under study), then the form of the bounds (3.16) becomes more or less similar to that in the monodisperse case. Namely, in the frame of this assumption

$$\mathcal{R}_0(\mathbf{x}_{12}; s_1, s_2) = P(s_1) P(s_2) R(\mathbf{x}_{12}), \quad \mathcal{F}_0(\mathbf{y}; s_1, s_2) = P(s_1) P(s_2) F_0(\mathbf{y}),$$

where

$$R(\mathbf{x}_{12}) = 1 - g_0(\mathbf{x}_{12}), \quad F_0(\mathbf{y}) = \int h(\mathbf{x}) R(\mathbf{x} - \mathbf{y})\, d^3\mathbf{x}$$

are the same functions that appeared in [8] when dealing with the monodisperse case. Then the formulae (3.16) simplify:

$$a^l_{2\kappa} = 3\left\{N^2 + \langle\beta^2(\widetilde{s})\rangle\left\langle\frac{K_f\widetilde{s}-\kappa_m}{K_f\widetilde{s}}\right\rangle m_2\right\},$$

$$a^u_{2\kappa} = 3\left\{N^2 + \langle\beta^2(\widetilde{s})\rangle\frac{K_f-\kappa_m}{\kappa_m}m_2\right\}, \tag{3.21}$$

where

$$m_2 = -\frac{1}{2\pi}\int_0^\infty \frac{F_0'(r)}{r^3}\,dr = 2\int_2^\infty \frac{\lambda^2}{(\lambda^2-1)^3}g_0(\lambda a)\,d\lambda \tag{3.22}$$

is the same statistical parameter as in the monodisperse case, see [8]. In particular, if the spheres have non-random conductivity $\kappa_f = K_f$, then

$$a^l_{2\kappa} = 3\beta^2\left\{1+\frac{[\kappa]}{\kappa_f}m_2\right\},\quad a^u_{2\kappa} = 3\beta^2\left\{1+\frac{[\kappa]}{\kappa_m}m_2\right\}, \tag{3.23}$$

where $[\kappa] = \kappa_f - \kappa_m$, $\beta = \beta(1) = [\kappa]/(\kappa_f + 2\kappa_m)$, see [8] again, which coincides with the monodisperse result of Markov [8]. Under the assumption (3.20) we shall consider the following two examples.

3.3.1. *"Triangular" distribution.* Since the conductivity $\widetilde{\kappa}_f = K_f\widetilde{s} \geq 0$, it is impossible to adopt the popular Gaussian distribution. That is way we consider the "triangular" (Simpson) distribution of $\widetilde{\kappa}_f$ in the interval $[K_1, K_2]$ as a certain counterpart of the Gaussian one. Then

$$P(s) = \begin{cases} \dfrac{2K_f}{K_2-K_1}\left[1-\dfrac{|K_1+K_2-2K_fs|}{K_2-K_1}\right] & \text{at} \quad K_fs\in[K_1,K_2], \\ 0 \qquad \text{otherwise,} \end{cases} \tag{3.24}$$

where $K_f = (K_1+K_2)/2$. After simple algebra, based of Eqs. (3.15) and (3.24), we get

$$a_{1\kappa} = 3N,\quad N = 1+\frac{6}{\gamma^2\omega^2}\Big[4(\gamma+2)\ln(2\gamma+4)$$

$$-(\gamma(2-\omega)+4)\ln(\gamma(2-\omega)+4)-(\gamma(2+\omega)+4)\ln(\gamma(2+\omega)+4)\Big], \tag{3.25}$$

where $\gamma = K_f/k_m$ and $\omega = (K_2-K_1)/K_f$ is, so to say, the "divergence" of the non-dimensional sphere conductivity. Since $K_2 \geq K_1 \geq 0$, then $\gamma \geq 0$ and $0 \leq \omega \leq 2$.

Similarly, with Eq. (3.20) taken into account, the bounds (3.16) read

$$a^l_{2\kappa} = 3\left\{N^2+\Upsilon(\gamma,\omega)\Lambda(\gamma,\omega)m_2\right\},\quad a^u_{2\kappa} = 3\left\{N^2+(\gamma-1)\Lambda(\gamma,\omega)m_2\right\}, \tag{3.26a}$$

where

$$\Upsilon(\gamma,\omega) = 1-\frac{2}{\gamma\omega^2}\Big[(2-\omega)\ln(2-\omega)+(2+\omega)\ln(2+\omega)-4\ln 2\Big], \tag{3.26b}$$

$$\Lambda(\gamma,\omega) = 1 + \frac{12}{\gamma^2\omega^2}\Big[2(2\gamma+7)\ln(2\gamma+4)$$
$$-(\gamma(2-\omega)+7)\ln(\gamma(2-\omega)+4) - (\gamma(2+\omega)+7)\ln(\gamma(2+\omega)+4)\Big]. \quad (3.26c)$$

The quantities $a_{1\kappa}$, $a^u_{1\kappa(B)}$, $a^l_{2\kappa}$ and $a^u_{2\kappa}$ as functions of the parameter ω are shown in Fig. 1 and 2 for $\gamma = 5$. The "well-stirred" case $g_0(r) = 1$ at $r > 2a$ is considered, when $m_2 = \frac{5}{18} - \frac{1}{8}\ln 2 \approx 0.14045$, see [8]. In Fig. 1 the value of approximation $\widetilde{a}_{1\kappa} = 3(\gamma-1)/(\gamma+2)$ for $a_{1\kappa}$ is also given, which corresponds to the rough assumption that the dispersion is replaced with a monodisperse one of

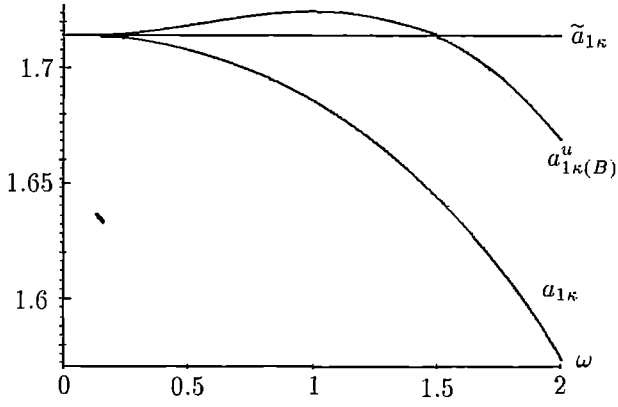


Fig. 1. The variations of the c-coefficient $a_{1\kappa}$ of the effective conductivity of the dispersion with "divergence" ω in the "triangular" case ($\gamma = K_f/\kappa_m = 5$); $a_{1\kappa}$ — the exact value (3.15); $a^u_{1\kappa(B)}$ — the Beran's upper bound, see (3.27); $\widetilde{a}_{1\kappa}$ — the "monodisperse" approximation

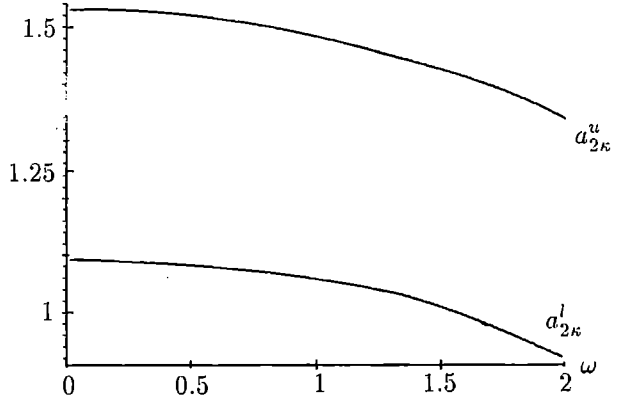


Fig. 2. The variations of the c^2-bounds $a^l_{2\kappa}$ and $a^u_{2\kappa}$ of the effective conductivity of the dispersion with "divergence" ω in the "triangular" case ($\gamma = K_f/\kappa_m = 5$)

sphere's conductivity that equals the mean value K_f (the "monodisperse" approximation). It is seen that this approximation is non-realistic; it is only justified at the limit case $\omega \to 0$. The dependence on ω of the upper Beran's bound $a^u_{1\kappa(B)}$, which now has the form

$$a^u_{1\kappa(B)} = \gamma - 1 - \frac{1}{72}\frac{(\gamma^2\omega^2 + 24(\gamma-1)^2)^2}{\gamma^3\omega^2 + 8(\gamma-1)^2(\gamma+2)}, \tag{3.27}$$

is plotted as well in Fig. 1.

3.3.2. *A Dispersion Containing Two Kinds of Spheres.* Consider the case when there exist only two kinds of spheres in the dispersion, having the conductivities $\kappa_f^{(1)}$, $\kappa_f^{(2)}$ and volume fractions c_1, c_2, respectively, $c = c_1 + c_2$. Then

$$P(s) = p_1\delta(s - s^{(1)}) + p_2\delta(s - s^{(2)}),$$

where $s^{(i)} = \kappa_f^{(i)}/K_f$, $p_i = c_i/c$, $i = 1, 2$, $K_f = p_1\kappa_f^{(1)} + p_2\kappa_2^{(2)}$. In this case the c-coefficient $a_{1\kappa}$ becomes

$$a_{1\kappa} = 3(p_1\beta_1 + p_2\beta_2),$$

where

$$\beta_i = \beta(s^{(i)}) = \frac{\kappa_f^{(i)} - \kappa_m}{\kappa_f^{(i)} + 2\kappa_m} = \frac{\alpha_i - 1}{\alpha_i + 2}, \quad \alpha_i = \frac{\kappa_f^{(i)}}{\kappa_m}, \quad i = 1, 2.$$

Similarly, the bounds (3.16) on the c^2-coefficient $a_{2\kappa}$ read

$$a^l_{2\kappa} = 3\sum_{i=1}^{2}\left(p_i\beta_i + \left(1 - \frac{p_1}{\alpha_1} - \frac{p_2}{\alpha_2}\right)\beta_i^2 m_2\right),$$

$$a^u_{2\kappa} = 3\sum_{i=1}^{2}\left(p_i\beta_i + (p_1\alpha_1 + p_2\alpha_2 - 1)\beta_i^2 m_2\right).$$

Let us note that the dispersion under study represents a three-phase medium: in the matrix two types of spherical particles of different conductivities are distributed. The generation of the above formulae for n-phase media of this kind is straightforward.

4. VARIATIONAL DERIVATION OF c^2-FORMULA FOR THE EFFECTIVE CONDUCTIVITY OF THE DISPERSION

Consider now the series (2.14) as a class of trial fields:

$$T_A^{(2)} = \left\{\theta(\mathbf{x}) \mid \theta(\mathbf{x}) = \mathbf{G}\cdot\mathbf{x} + \int\!\!\int T_1(\mathbf{x}-\mathbf{y}, s)D_\psi^{(1)}(\mathbf{y}; s)\, d^3\mathbf{y}\, ds\right.$$

$$\left. + \int\!\!\int\!\!\int\!\!\int T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2)D_\psi^{(2)}(\mathbf{y}_1, \mathbf{y}_2; s_1, s_2)\, d^3\mathbf{y}_1\, d^3\mathbf{y}_2\, ds_1\, ds_2\right\}, \tag{4.1}$$

where now the kernels $T_1(\mathbf{x}, s)$ and $T_2(\mathbf{x}, \mathbf{y}, s_1, s_2)$ are adjustable. Using the formulae for the moments of the fields $D_\psi^{(1)}$ and $D_\psi^{(2)}$, see Eqs. (2.15), the restriction $W_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)]$ of the functional W_A over this class becomes

$$W_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)] = W_A\big|_{\mathcal{T}_A^{(2)}} = W_A^{(1)}[T_1(\cdot)] + \widetilde{W}_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)],$$

where

$$\widetilde{W}_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)] = n^2\kappa_m \iiint g_0(\mathbf{y}_{2,1}) P_2(s_1, s_2) \Big[|\nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2)|^2$$

$$+ \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2) \cdot \nabla_x T_2(\mathbf{x}-\mathbf{y}_2, \mathbf{x}-\mathbf{y}_1, s_2, s_1) \Big] \, d^3\mathbf{y}_1 \, d^3\mathbf{y}_2 \, ds_1 \, ds_2$$

$$+ 2n^2 \iiint g_0(\mathbf{y}_{2,1}) P_2(s_1, s_2) \Big[(K_f s_1 - \kappa_m) h(\mathbf{x}-\mathbf{y}_1) \nabla T_1(\mathbf{x}-\mathbf{y}_2, s_2)$$

$$+ (K_f s_2 - \kappa_m) h(\mathbf{x}-\mathbf{y}_2) \nabla T_1(\mathbf{x}-\mathbf{y}_1, s_1) \Big] \cdot \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2) \, d^3\mathbf{y}_1 \, d^3\mathbf{y}_2 \, ds_1 \, ds_2$$

$$+ n^2 \iiint g_0(\mathbf{y}_{2,1}) P_2(s_1, s_2) \big[(K_f s_1 - \kappa_m) h(\mathbf{x}-\mathbf{y}_1) + (K_f s_2 - \kappa_m) h(\mathbf{x}-\mathbf{y}_2) \big]$$

$$\times \big[|\nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2)|^2 + \nabla_x T_2(\mathbf{x}-\mathbf{y}_1, \mathbf{x}-\mathbf{y}_2, s_1, s_2)$$

$$\cdot \nabla_x T_2(\mathbf{x}-\mathbf{y}_2, \mathbf{x}-\mathbf{y}_1, s_2, s_1) \big] \, d^3\mathbf{y}_1 \, d^3\mathbf{y}_2 \, ds_1 \, ds_2 + o(n^2).$$

The optimal kernels $T_1(\mathbf{x}, s)$ and $T_2(\mathbf{x}, \mathbf{y}, s_1, s_2)$ are looked for again in the virial form (3.5) for T_1 and

$$T_2(\mathbf{x}, \mathbf{y}, s_1, s_2) = T_2(\mathbf{x}, \mathbf{y}, s_1, s_2; n)$$

$$= T_{2,0}(\mathbf{x}, \mathbf{y}, s_1, s_2) + T_{2,1}(\mathbf{x}, \mathbf{y}, s_1, s_2)\, n + \cdots$$

for T_2, which implies the respective virial expansion of the functional $W_A^{(2)}$, namely,

$$W_A^{(2)}[T_1(\cdot), T_2(\cdot,\cdot)] = \langle\kappa\rangle G^2 + W_A^{(1,1)}[T_{1,0}(\cdot)]\, n$$

$$+ W_A^{(2,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot), T_{2,0}(\cdot,\cdot)]\, n^2 + o(n^2), \tag{4.2}$$

where

$$W_A^{(2,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot), T_{2,0}(\cdot,\cdot)]$$

$$= W_A^{(1,2)}[T_{1,0}(\cdot), T_{1,1}(\cdot)] + \widetilde{W}_A^{(2)}[T_{1,0}(\cdot), T_{2,0}(\cdot,\cdot)]; \tag{4.3}$$

here $W_A^{(1,1)}$ and $W_A^{(1,2)}$ are the virial coefficients from Eq. (3.6) for which, let us recall, Eqs. (3.11) and (3.12) hold. Hence, the minimization of the functional $W_A^{(2)}$ is reduced to that of the functional

$$\widetilde{W}_A^{(2)\dagger}[T_{2,0}(\cdot,\cdot)] = \widetilde{W}_A^{(2)}[T^{(1)}(\cdot), T_{2,0}(\cdot,\cdot)].$$

The Euler-Lagrange equation for the latter is

$$P_2(s_1,s_2)\Big\{\kappa_m(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})\cdot\Big[g_0(\mathbf{z}_1-\mathbf{z}_2)(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})\widetilde{T}_{2,0}(\mathbf{z}_1,\mathbf{z}_2,s_1,s_2)\Big]$$

$$+(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})\cdot\Big[g_0(\mathbf{z}_1-\mathbf{z}_2)\big[(K_f s_1-\kappa_m)h(\mathbf{z}_1)\nabla T^{(1)}(\mathbf{z}_2,s_2)$$

$$+(K_f s_2-\kappa_m)h(\mathbf{z}_2)\nabla T^{(1)}(\mathbf{z}_1,s_1)\big]\Big]$$

$$+(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})\cdot\Big[g_0(\mathbf{z}_1-\mathbf{z}_2)\big[(K_f s_1-\kappa_m)h(\mathbf{z}_1)+(K_f s_2-\kappa_m)h(\mathbf{z}_2)\big]$$

$$+(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})\widetilde{T}_{2,0}(\mathbf{z}_1,\mathbf{z}_2,s_1,s_2)\Big]\Big\}=0 \tag{4.4}$$

with the notation

$$\widetilde{T}_{2,0}(\mathbf{z}_1,\mathbf{z}_2,s_1,s_2)=T_{2,0}(\mathbf{z}_1,\mathbf{z}_2,s_1,s_2)+T_{2,0}(\mathbf{z}_2,\mathbf{z}_1,s_2,s_1).$$

Taking into account that $(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})g_0(\mathbf{z}_1-\mathbf{z}_2)=0$, an appropriate change of variables allows to recast Eq. (4.4) as

$$g_0(\mathbf{z})P_2(s_1,s_2)\Big\{\kappa_m\Delta_x\widetilde{T}_{2,0}(\mathbf{x},\mathbf{x}-\mathbf{z},s_1,s_2)$$

$$+\nabla_x\cdot\big[(K_f s_1-\kappa_m)h(\mathbf{x})\nabla T^{(1)}(\mathbf{x}-\mathbf{z},s_2)+(K_f s_2-\kappa_m)h(\mathbf{x}-\mathbf{z})\nabla T^{(1)}(\mathbf{x},s_1)\big] \tag{4.5}$$

$$+\nabla_x\cdot\Big[\big[(K_f s_1-\kappa_m)h(\mathbf{x})+(K_f s_2-\kappa_m)h(\mathbf{x}-\mathbf{z})\big]\nabla_x\widetilde{T}_{2,0}(\mathbf{x},\mathbf{x}-\mathbf{z},s_1,s_2)\Big]\Big\}=0.$$

Similarly to the monodisperse case [5], the solution of Eq. (4.5) is

$$\widetilde{T}_{2,0}(\mathbf{x},\mathbf{x}-\mathbf{z},s_1,s_2)=T^{(2)}(\mathbf{x},s_1;\mathbf{z},s_2)-T^{(1)}(\mathbf{x},s_1)-T^{(1)}(\mathbf{x}-\mathbf{z},s_2)\,, \tag{4.6}$$

where $T^{(2)}(\mathbf{x},s_1;\mathbf{y},s_2)$ is the disturbance to the temperature field $\mathbf{G}\cdot\mathbf{x}$ in an unbounded matrix of conductivity κ_m, generated by two spherical inhomogeneities: one of conductivity $K_f s_1$ located at the origin, and the other of conductivity $K_f s_2$ located at the point $\mathbf{y}$.

Making use of Eq. (4.4), the minimum value of the functional $\widetilde{W}_A^{(2)\dagger}$ can be recast now in the form in which the field $T_{2,0}(\mathbf{x},\mathbf{y},s_1,s_2)$ enters linearly:

$$\min\widetilde{W}_A^{(2)\dagger}[T_{2,0}(\cdot,\cdot)]=n^2\iiiint g_0(\mathbf{z}_1-\mathbf{z}_2)P_2(s_1,s_2)$$

$$\times\big[(K_f s_1-\kappa_m)h(\mathbf{z}_1)\nabla T^{(1)}(\mathbf{z}_2,s_2)+(K_f s_2-\kappa_m)h(\mathbf{z}_2)\nabla T^{(1)}(\mathbf{z}_1,s_1)\big]$$

$$\cdot(\nabla_{\mathbf{z}_1}+\nabla_{\mathbf{z}_2})T_{2,0}(\mathbf{z}_1,\mathbf{z}_2,s_1,s_2)\,d^3\mathbf{z}_1\,d^3\mathbf{z}_2\,ds_1\,ds_2$$

$$=n^2\iiiint P_2(s_1,s_2)(K_f s_1-\kappa_m)g_0(\mathbf{y})h(\mathbf{x})\nabla T^{(1)}(\mathbf{x}-\mathbf{y},s_2)$$

$$\cdot\big[\nabla_x T^{(2)}(\mathbf{x},s_1;\mathbf{y},s_2)-\nabla T^{(1)}(\mathbf{x},s_1)-\nabla T^{(1)}(\mathbf{x}-\mathbf{y},s_2)\big]\,d^3\mathbf{x}\,d^3\mathbf{y}\,ds_1\,ds_2. \tag{4.7}$$

Taking into account Eqs. (4.2), (4.3), (3.11), (3.12) and the formulae

$$\int h(\mathbf{x})\, d^3\mathbf{x} \int g_0(\mathbf{y}) \nabla T^{(1)}(\mathbf{x}-\mathbf{y}, s_2) \cdot \nabla T^{(1)}(\mathbf{x}, s_1)\, d^3\mathbf{y} = 0,$$

$$P_2(s_1, s_2) \int h(\mathbf{x})\, d^3\mathbf{x} \int g_0(\mathbf{y}) |\nabla T^{(1)}(\mathbf{x}-\mathbf{y}, s_2)|^2\, d^3\mathbf{y} = 3\beta^2(s_2) \mathcal{M}_2(s_1, s_2) V_a^2,$$

one finds for the c^2-coefficient

$$a_{2\kappa} = 3N^2 + a'_{2\kappa},$$

where

$$a'_{2\kappa} = \frac{1}{V_a^2} \iint P_2(s_1, s_2) \frac{K_f s_1 - \kappa_m}{\kappa_m}\, ds_1 ds_2$$

$$\times \int h(\mathbf{x})\, d^3\mathbf{x} \int g_0(\mathbf{y}) \nabla_x T^{(1)}(\mathbf{x}-\mathbf{y}, s_2) \cdot \nabla_x T^{(2)}(\mathbf{x}, s_1; \mathbf{y}, s_2)\, d^3\mathbf{y}. \tag{4.8}$$

Let us recall that the latter result follows from the fact that the solution of the random problem (1.2), asymptotically valid to the order c^2, is one of the trial fields from the class $\mathcal{T}_A^{(2)}$, see Sec. 2.2, over which the energy functional $W_A[\theta(\cdot)]$ is minimized. The formula (4.8) is the counterpart of the formula (3.9) in [14] in the monodisperse case. Note that the formula (4.8) contains absolutely convergent integrals only, see [10, 11] for details, so that no "renormalization" is needed, similar to that used by Jeffrey [19].

Finally, it is to be noted that the coefficient $T_{1,1}(\mathbf{x}, s)$ in the expansion (3.5) cannot be found within the frame of the above performed variational n^2-analysis. For the full solution of the random problem (1.2) to the order c^2 in the explained above sense it is necessary, however, to know the virial coefficients $T_{1,0}(\mathbf{x}, s)$, $T_{1,1}(\mathbf{x}, s)$ and $T_{2,0}(\mathbf{x}, \mathbf{y}, s_1, s_2)$: for example, when evaluating the two-point correlation function $\langle \theta'(\mathbf{x}) \theta'(\mathbf{y}) \rangle$, the convolution $\int T_{1,0}(\mathbf{x}-\mathbf{y}, s) T_{1,1}(\mathbf{y}, s)\, d^3\mathbf{y}$ appears, see [6] for details. That is why, in order to obtain function $T_{1,1}(\mathbf{x}, s)$ and as a consequence the full statistical solution of problem (1.2) to the order c^2, either the higher degrees of n in the virial expansion of the functional W_A should be considered or the procedure of Christov and Markov [5] should be employed instead. In the latter, however, conditionally convergent integral in the formula for the effective conductivity will show up with a correct mode of integration extracted in the course of the appropriate solution, see again [2, 6, 7] for details.

ACKNOWLEDGEMENTS. The support of this work by the Bulgarian Ministry of Education, Science and Technology under Grant No MM416-94 is gratefully acknowledged. The author thanks K. Z. Markov for helpful and stimulating discussions.

REFERENCES

1. Christov, C. I., K. Z. Markov. Stochastic functional expansion for heat conductivity of polydisperse perfectly disordered suspensions. *Annuaire Univ. Sofia, Fac. Math. Méc., Livre 2*, **79**, 1988, 191–207.
2. Markov, K. Z., C. I. Christov. On the problem of heat conduction for random dispersions of spheres allowed to overlap. *Math. Models and Methods in Applied Sciences*, **2**, 1992, 249–269.
3. Snyder, D. Random point processes. John Wiley, New York, 1975.
4. Beran, M. Statistical continuum theories. John Wiley, New York, 1968.
5. Christov, C. I., K. Z. Markov. Stochastic functional expansion for random media with perfectly disordered constitution. *SIAM J. Appl. Math.*, **45**, 1985, 289–311.
6. Markov, K. Z. On the heat propagation problem for random dispersions of spheres. *Math. Balkanica (New Series)*, **3**, 1989, 399–417.
7. Markov, K. Z. On the factorial functional series and their application to random media. *SIAM J. Appl. Math.*, **51**, 1991, 172–186.
8. Markov, K. Z. Application of Volterra-Wiener series for bounding the overall conductivity of heterogeneous media. I. General procedure. II. Suspensions of equi-sized spheres. — *SIAM J. Appl. Math.*, **47**, 1987, 831–850, 851–870.
9. Markov, K. Z., K. D. Zvyatkov. Optimal third-order bounds on the effective properties of some composite media, and related problems. *Advances in Mechanics (Warsaw)*, **14**, No 4, 1991, 3–46.
10. Markov, K. Z., K. D. Zvyatkov. Functional series and Hashin-Shtrikman's type bounds on the effective conductivity of random media. *Europ. J. Appl. Math.*, **6**, 1995, 611–629.
11. Markov, K. Z., K. D. Zvyatkov. Functional series and Hashin-Shtrikman's type bounds on the effective properties of random media. In: *Advances in Mathematical Modeling of Composite Materials*, ed. K. Z. Markov, World Sci., 1994, 59-106.
12. Markov, K. Z., K. D. Zvyatkov. On the optimal third-order bounds on the effective conductivity of random dispersions of spheres, *J. Theor. Appl. Mech.*, Bulg. Acad. Sci., **22**(3), 1991, 107–116.
13. Beran, M. Use of a variational approach to determine bounds for the effective permittivity of a random medium. *Nuovo Cimento*, **38**, 1965, 771–782.
14. Zvyatkov, K. D. Variational principles and the c^2-formula for the effective conductivity of a random dispersion. In: *Continuum Models and Discrete Systems*, ed. K. Z. Markov, World Sci., 1996, 324–331.
15. Stratonovich, R. L. Topics in theory of random noises, Vol. 1, Gordon and Breach, New York, 1967.
16. Torquato, S. Bulk properties of two-phase disordered media. III. New bounds on the effective conductivity of dispersions of penetrable spheres. *J. Chem. Phys.*, **84**, 1986, 6345–6359.
17. Zvyatkov, K. D. On the effective properties of a random composite medium. In: *Proceedings of 6^{th} Bulg. Congress on Mechanics*, 1989, Varna, Bulgaria, vol. 2, 1990, 240–243 (in Bulgarian).
18. Zvyatkov, K. D. On the effective properties of polydisperse composite materials. In: *Proceedings of International Youth School "Application of mechanics in robotics and new materials"*, Sunny Beach, Bulgaria, 1988, 322–326.
19. Jeffrey, D. J. Conduction through a random suspension of spheres. *Proc. Roy. Soc. London*, **A335**, 1973, 355–367.

Received on September 15, 1996
Revised on October 22, 1997

Faculty of Mathematics and Informatics
"K. Preslavski" University of Shumen
BG-9700 Shumen, Bulgatia
e-mail: zvjatkov@uni-shoumen.bg

Submission of manuscripts. The *Annuaire* is published once a year, in two parts: Part I. Mathematics and Mechanics, and Part II. Applied Mathematics and Informatics. No deadline exists. Once received by the editors, the manuscript will be subjected to rapid, but thorough review process. If accepted, it is immediately sheduled for the nearest forthcoming issue. No page charge is made. The author(s) will be provided with a total of 30 free of charge offprints of their paper.

The submission of a paper implies that it has not been published, or is not under consideration for publication elsewhere. In case it is accepted, it implies as well that the author(s) transfers the copyright to the Faculty of Mathematics and Informatics at the "St. Kliment Ohridski" University of Sofia, including the right to adapt the article for use in conjuction with computer systems and programs and also reproduction or publication in machine-readable form and incorporation in retrieval systems.

Instructions to Contributors. Preferences will be given to papers, not longer than 15 to 20 pages, written preferably in English and typeset by means of a TEX system. A simple specimen file, exposing in detail the instruction for preparing the manuscripts, is available upon request from the electronic address of the Editorial Board. Two copies of the manuscript should be submitted. Upon acceptance of the paper, the authors will be asked to send by electronic mail or on a diskette the text of the papers and the appropriate graphic files (in any format like `*.tif`, `*.pcx`, `*.bmp`, etc.).

The manuscripts should be prepared for publication in accordance with the instructions, given below.

The manuscripts must be *typed* on one side of the paper in double spacing with wide margins. On the *first* page the author should provide: a title, name(s) of the author(s), a short abstract, a list of keywords and the appropriate 1995 Mathematical Subject Classification codes (primary and secondary, if necessary). The affiliation(s), including the electronic address, is given at the end of the manuscripts.

Figures have to be inserted in the text near their first reference. If the author cannot supply and/or incorporate the graphic files, drawings (in black ink and on a good quality paper) should be enclosed separately. If photographs are to be used, only black and white ones are acceptable.

Tables should be inserted in the text as close to the point of reference as possible. Some space should be left above and below the table.

Footnotes, which should be kept to a minimum and should be brief, must be numbered consecutively.

References must be cited in the text in square brackets, like [3], or [5, 7], or [11, p. 123], or [16, Ch. 2.12]. They have to be numbered either in the order they appear in the text or alphabetically. Examples (please note order, style and punctuation):

For books: Obreshkoff, N. Higher algebra. Nauka i Izkustvo, 2nd edition, Sofia, 1963 (in Bulgarian).

For journal articles: Frisch, H. L. Statistics of random media. *Trans. Soc. Rheology*, 9, 1965, 293–312.

For articles in edited volumes or proceedings: Friedman, H. Axiomatic recursive function theory. In: *Logic Colloquim 95*, eds. R. Gandy and F. Yates, North-Holland, 1971, 188–195.