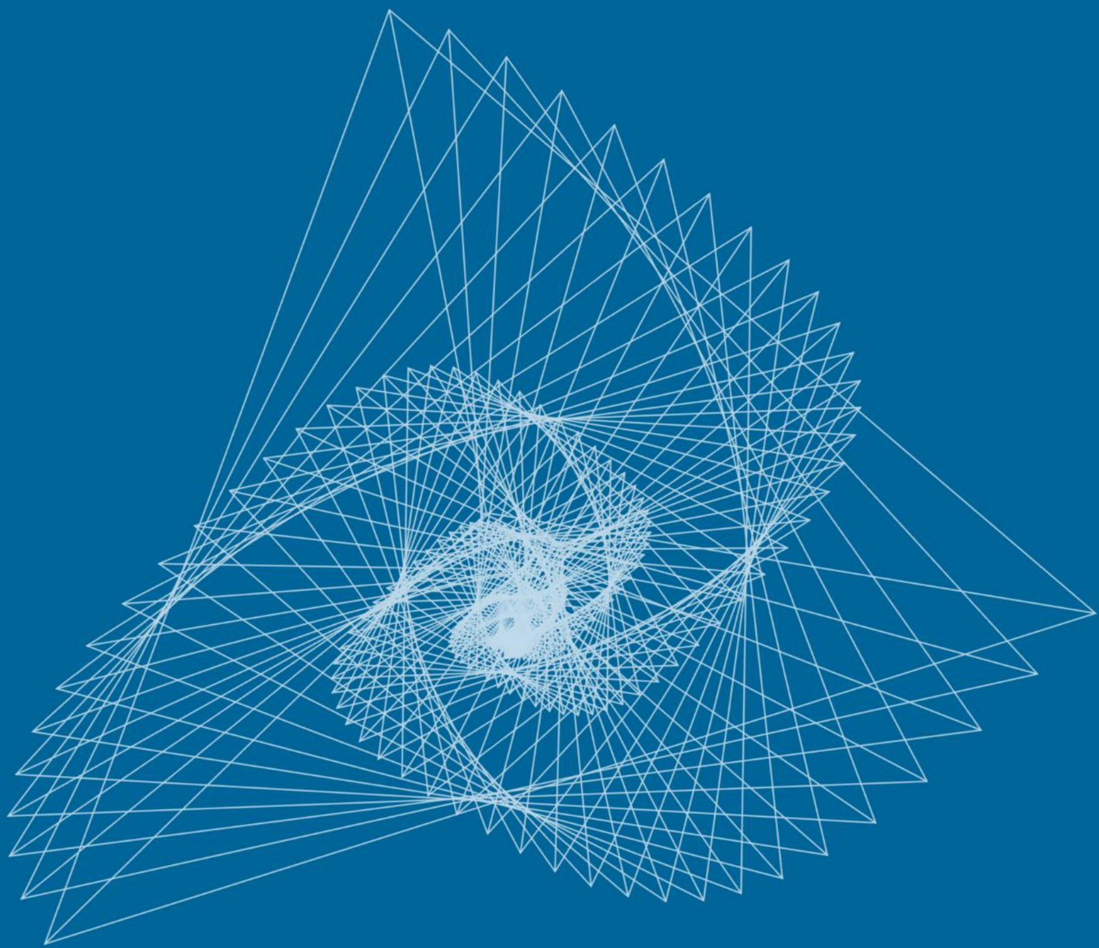
VOL. 4, NO. 2

URAL MATHEMATICAL JOURNAL

N.N. Krasovskii Institute of Mathematics and Mechanics of
the Ural Branch of Russian Academy of Sciences and
Ural Federal University named after the first President of Russia B.N.Yeltsin

ISSN: 2414-3952



2018

ISSN 2414-3952
(Online)

Ural Mathematical Journal

Vol. 4, no. 2, 2018

https://umjuran.ru

Electronic Periodical Scientific Journal
Founded in 2015

The Journal is registered by the Federal Service for Supervision in the Sphere of Communication, Information Technologies and Mass Communications
Certificate of Registration of the Mass Media Эл № ФС77-61719 of 07.05.2015

Founders

N.N.Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N.Yeltsin

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990
Phone: +7 (343) 375-34-73   Fax: +7 (343) 374-25-81
Email: secretary@umjuran.ru
Web-site: https://umjuran.ru

TABLE OF CONTENTS

EXPERT REVIEWS AND OPINIONS

PART I. PROCEEDINGS OF THE 43RD S.B. STECHKIN'S WORKSHOP

PART II. GENERAL TOPICS

LETTERS TO THE EDITORIAL BOARD

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 3–5

DOI: 10.15826/umj.2018.2.001

ON THE 75TH BIRTHDAY OF PROFESSOR VITALII VLADIMIROVICH ARESTOV

Vitalii I. Berdyshev[†], **Nikolai I. Chernykh**[††], **Alexander G. Babenko**[†††], **and Roman R. Akopyan**[††††]

Krasovskii Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences, 16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990;

Ural Federal University, 51 Lenin aven., Ekaterinburg, Russia, 620000

[†]bvi@imm.uran.ru, [††]chernykh@imm.uran.ru, [†††] babenko@imm.uran.ru, [††††]rrakopyan@mephi.ru

Figure 1. Professor Vitalii Vladimirovich Arestov

July 16, 2018, was the 75th birthday of famous Russian scientist, prominent mathematician, Doctor of Physics and Mathematics, Professor Vitalii Vladimirovich Arestov.

Arestov was born in the village Bol'shoi Karai (Saratov oblast). He graduated from Saratov State University as specialist in mathematics and completed postgraduate studies in Steklov Mathematical Institute of the USSR Academy of Sciences under supervision of Professor S.B. Stechkin. Starting from 1968, Vitalii Vladimirovich works at the Sverdlovsk Division of the Steklov Mathematical Institute of the USSR Academy of Sciences (now the Krasovskii Institute of Mathematics and Mechanics, Ural Branch of Russian Academy of Sciences (IMM UB RAS)). Starting from 1970, he also teaches at Ural State University (now Ural Federal University (UrFU)), where he headed the Department of Mathematical Analysis and Function Theory (now Department of Mathematical Analysis). At present, Arestov is a professor at UrFU and a leading researcher at the Department of Approximation and Applications of IMM UB RAS.

Arestov is a prominent specialist in function theory. He has profound fundamental results on the best approximation of operators by operators of a simpler structure, in particular, on the best approximation of unbounded operators by bounded ones and on related problems on the smallest constants in inequalities between norms of derivatives of differentiable functions, on the best approximation of one class of differentiable functions by another class of smoother functions, and optimal recovery of unbounded operators on classes of elements given with error. Recently, in this topic, he solved the problem on the best uniform approximation on the numerical axis of the first-order differentiation operator on the class of functions with a bounded second derivative by bounded linear operators acting from the space of functions with a summable Fourier transform to the space of continuous functions. He also obtained the related exact Kolmogorov-type inequality between the uniform norm of the derivative of a function, the L-norm of the Fourier transform of the function, and the L_∞-norm of its second derivative. In addition, Arestov and M.A. Filatova found a solution to the long standing problem of the best approximation of the differentiation operator by linear bounded operators on the class of twice differentiable functions in the space L_2 on the semi-axis. He obtained important results on the extremal properties of polynomials; in particular, an amazing result widely known to specialists in the theory of extremal problems about the best constant in the Bernstein inequality for trigonometric polynomials in the space L_p for $0 \le p < 1$. A method has been created for studying extremal problems for polynomials in the space L_p and in the more general case for Orlicz type φ-spaces. Fundamental results were obtained in extremal problems for positive definite functions, including problems related to the exact Jackson inequalities for the best approximations of functions by polynomials and entire functions as well as problems of Delsarte and Turan, which are related to discrete extremal geometric problems (some of these results were obtained together with Arestov's students A.G. Babenko and E.E. Berdysheva). In recent years, in a series of papers (together with M.V. Deikalova and others), a new approach, based on the properties of the generalized translation operator, has been developed to the study of extremal problems for polynomials and entire functions; subtle properties of generalized translation operators in spaces with different weights have been thoroughly investigated.

Arestov is the author of 85 scientific papers and several textbooks. The list of his main publications can be found at http://work.imkn.urfu.ru/arestov/index_en.html.

Arestov is one of the leading lecturers of UrFU in continuous mathematics. His department provides a high level of training of specialists in function theory, contributing to the development of scientific cooperation between UrFU and IMM UB RAS. His perceptive and comprehensible lectures form an interest in mathematics; the diverse and relevant topics of special courses attract students, many of whom continue to conduct research under Arestov's supervision. Eleven of his students successfully defended their dissertations, one of them has already become a Doctor of Physics and Mathematics, and another one holds Habilitation degree in Germany.

Arestov was a member of the editorial boards of leading scientific journals *Mathematical Notes* and *East Journal on Approximations*. He is currently the deputy editor-in-chief of *Ural Mathematical Journal*. Arestov took an active part in organizing and conducting a number of high-level international conferences as co-chairman or member of their organizing committees and program committees. Two of them are regular: triennial International Conference *Algorithmic Analysis of Unstable Problems* (1995–2014) and annual *International S.B. Stechkin's School-Conference on Function Theory* (from 1971 to the present).

August 1–10, 2018, the 43d International S.B. Stechkin's School-Conference on Function Theory dedicated to the 75th birthday of Professor Vitalii Vladimirovich Arestov was held in Kyshtym, Chelyabinskaya oblast. It was organized by UrFU and IMM UB RAS.

The collectives of Ural Federal University and the Institute of Mathematics and Mechanics, the editorial board of *Ural Mathematical Journal*, friends, colleagues, and students, heartily congratulate Vitalii Vladimirovich on his glorious jubilee and wish him further success in scientific, educational, and social activities, good health, and great personal happiness.

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 6–12
DOI: 10.15826/umj.2018.2.002

ON THE SUMMABILITY OF THE DISCRETE HILBERT TRANSFORM

Rashid A. Aliev

Baku State University, Baku, AZ 1148, Azerbaijan;
Institute of Mathematics and Mechanics, NAS of Azerbaijan, Baku, AZ 1141, Azerbaijan
aliyevrashid@mail.ru

Aynur F. Amrahova

Baku State University, Baku, AZ 1148, Azerbaijan
amrahovaaynur@mail.ru

Abstract: In this paper, we study the asymptotic behavior of the distribution function of the discrete Hilbert transform of sequences from the class l_1 and find a necessary condition and a sufficient condition for the summability of the discrete Hilbert transform of a sequence from the class l_1.

Keywords: Discrete Hilbert transform, Asymptotic behavior of the distribution function, Class of summable sequences.

Introduction

Denote by l_p, $p \geq 1$, the class of numeric sequences $b = \{b_n\}_{n \in Z}$ satisfying the condition

$$\|b\|_{l_p} = \Big(\sum_{n \in Z} |b_n|^p \Big)^{1/p} < \infty,$$

where Z is the set of integers.

Let $b = \{b_n\}_{n \in Z} \in l_1$. The sequence $H(b) = \{(Hb)_n\}_{n \in Z}$ is called the Hilbert transform of the sequence $b = \{b_n\}_{n \in Z}$, where

$$(Hb)_n = \sum_{m \neq n} \frac{b_m}{n - m}, \quad n \in Z.$$

M. Riesz proved (see [10] and [4, 7]) that, if $b \in l_p$, $p > 1$, then $H(b) \in l_p$ and the inequality

$$\|H(b)\|_{l_p} \leq C_p \|b\|_{l_p} \tag{0.1}$$

holds. Weighted analogues of (0.1) were investigated in [1–3, 5, 6, 8, 9, 11].

If $b \in l_1$, then the sequence $H(b)$ belongs to the class $\bigcap_{p>1} l_p$ but doesn't belong to the class l_1. In this case, R. Hunt, B. Muckenhoupt, and R. Wheeden proved (see [6]) that the distribution function

$$(Hb)(\lambda) \equiv \sum_{\{n \in Z:\ |(Hb)_n| > \lambda\}} 1$$

of the Hilbert transform of the sequence b satisfies the condition

$$\forall \lambda > 0 \quad |(Hb)(\lambda)| \leq \frac{C_0}{\lambda} \|b\|_{l_1}, \tag{0.2}$$

where C_0 is an absolute constant.

In this paper, we study the asymptotic behavior of the distribution function $(Hb)(\lambda)$ of the Hilbert transform of a sequence $b \in l_1$ as $\lambda \to 0$ and find a necessary condition and a sufficient condition for the summability of the discrete Hilbert transform of a sequence from the class l_1.

1. Asymptotic behavior of the distribution function of the discrete Hilbert transform

Theorem 1. *Let $b \in l_1$. Then the following equation holds*:

$$\lim_{\lambda \to 0+} \lambda \cdot (Hb)(\lambda) = 2\Big|\sum_{n \in Z} b_n\Big|. \tag{1.1}$$

We first prove an auxiliary lemma.

Lemma 1. *Let $b \in l_1$ and $\sum_{n \in Z} b_n = 0$. Then the following equation holds*:

$$(Hb)(\lambda) = o(1/\lambda), \quad \lambda \to 0+. \tag{1.2}$$

P r o o f. Assume first that the sequence $b \in l_1$ is concentrated on some finite interval $[-m, m]$, i. e., $b_n = 0$ for $|n| > m$. In this case, from the equality

$$(Hb)_n = \sum_{|k| \leq m} \frac{b_k}{n-k} - \frac{1}{n-1/2} \sum_{|k| \leq m} b_k = \sum_{|k| \leq m} \frac{k - 1/2}{(n-k)(n-1/2)} b_k, \quad |n| > m$$

we get that

$$|(Hb)_n| \leq \frac{4}{n^2} \sum_{|k| \leq m} (k - 1/2)\, b_k$$

for large values of n, whence the asymptotic equation (1.2) follows.

Let us now consider the general case. From the condition $\sum_{n \in Z} b_n = 0$, it follows that, for all $\varepsilon > 0$ there exist sequences $b' = \{b'_n\}_{n \in Z} \in l_1$ and $b'' = \{b''_n\}_{n \in Z} \in l_1$ satisfying the condition $b = b' + b''$, where the sequence $b' \in l_1$ is concentrated on some finite interval $[-m, m]$ and $\sum_{n \in Z} b'_n = 0$, and the sequence $b'' \in l_1$ satisfies the inequality $\sum_{n \in Z} |b''_n| < \varepsilon/(4C_0)$, with the constant C_0 from (0.2). Since the sequence $b' \in l_1$ is concentrated on $[-m, m]$ and $\sum_{n \in Z} b'_n = 0$, equation (1.2) is satisfied for the sequence $b' \in l_1$, and, therefore, there exists $\lambda(\varepsilon) > 0$ such that the inequality

$$\lambda \left(Hb'\right)(\lambda/2) < \varepsilon/2 \tag{1.3}$$

holds for $0 < \lambda < \lambda(\varepsilon)$, where $(Hb')(\lambda) = \sum_{\{n \in Z:\ |(Hb')_n| > \lambda\}} 1$. On the other hand, inequality (0.2) implies that

$$\lambda \left(Hb''\right)(\lambda/2) \leq 2C_0 \sum_{n \in Z} |b''_n| < \varepsilon/2 \tag{1.4}$$

for all $\lambda > 0$, where $(Hb'')(\lambda) = \sum_{\{n \in Z:\ |(Hb'')_n| > \lambda\}} 1$. From inequalities (1.3) and (1.4) and the inclusion

$$\{n \in Z:\ |(Hb)_n| > \lambda\} \subset \{n \in Z:\ |(Hb')_n| > \lambda/2\} \bigcup \{n \in Z:\ |(Hb'')_n| > \lambda/2\}$$

we obtain that

$$\lambda \cdot (Hb)(\lambda) \le \lambda \left(Hb'\right)(\lambda/2) + \lambda \left(Hb''\right)(\lambda/2) < \varepsilon$$

for $0 < \lambda < \lambda(\varepsilon)$. This shows that equality (1.2) holds for all $b \in l_1$ satisfying the condition $\sum_{n\in Z} b_n = 0$. This completes the proof of Lemma 1. □

P r o o f of Theorem 1. In the case $\sum_{n\in Z} b_n = 0$, the statement of the theorem follows from Lemma 1. Consider the case $\sum_{n\in Z} b_n = \alpha \neq 0$. We use the following notation: $b'_n = b_n$ for $n \neq 0$, $b'_0 = b_0 - \alpha$, $b''_n = 0$ for $n \neq 0$, and $b''_0 = \alpha$. Then $b = b' + b''$, where $b' = \{b'_n\}_{n\in Z} \in l_1$ and $b'' = \{b''_n\}_{n\in Z} \in l_1$. Since $\sum_{n\in Z} b'_n = 0$, we obtain from Lemma 1 that

$$(Hb')(\lambda) = o\left(1/\lambda\right), \quad \lambda \to 0+. \tag{1.5}$$

Since $(Hb''t)_n = \alpha/n$ for $n \neq 0$ and $(Hb'')_0 = 0$, we have

$$(Hb'')(\lambda) \sim \frac{2\left|\alpha\right|}{\lambda}, \quad \lambda \to 0+. \tag{1.6}$$

For all $0 < \varepsilon < 1$, by the inclusions

$$\left\{n \in Z:\ \left|\left(Hb''\right)_n\right| > (1+\varepsilon)\lambda\right\} \setminus \left\{n \in Z:\ \left|\left(Hb'\right)_n\right| > \varepsilon\lambda\right\} \subset$$
$$\subset \left\{n \in Z:\ \left|(Hb)_n\right| > \lambda\right\} \subset$$
$$\subset \left\{n \in Z:\ \left|\left(Hb'\right)_n\right| > \varepsilon\lambda\right\} \bigcup \left\{n \in Z:\ \left|\left(Hb''\right)_n\right| > (1-\varepsilon)\lambda\right\}$$

and relations (1.5) and (1.6), we have

$$\frac{2\left|\alpha\right|}{1+\varepsilon} \le \liminf_{\lambda\to 0+} \lambda \cdot (Hb)(\lambda) \le \limsup_{\lambda\to 0+} \lambda \cdot (Hb)(\lambda) \le \frac{2\left|\alpha\right|}{1-\varepsilon}.$$

This implies equation (1.1) and completes the proof of Theorem 1. □

2. A necessary condition and a sufficient condition for the summability of the discrete Hilbert transform

Theorem 2. *Let $b \in l_1$. If $Hb \in l_1$, then it is necessary that the following equation holds:*

$$\sum_{n\in Z} b_n = 0. \tag{2.1}$$

P r o o f. We first we prove that, if $h = \{h_n\}_{n\in Z} \in l_1$, then the distribution function $h(\lambda) = \sum_{\{n\in Z:\ |h_n|>\lambda\}} 1$ of the sequence h satisfies the condition

$$h(\lambda) = o\left(1/\lambda\right), \quad \lambda \to 0+. \tag{2.2}$$

Note that the condition $h = \{h_n\}_{n\in Z} \in l_1$ implies that the set of $\{n \in Z:\ |h_n| > \lambda\}$ is finite for all $\lambda > 0$. Then, the inequality

$$\sum_{n\in Z} |h_n| = \sum_{\{n\in Z:\, |h_n|>1\}} |h_n| + \sum_{k=0}^{\infty} \left[\sum_{\left\{n\in Z:\, |h_n|\in\left(2^{-k-1};\, 2^{-k}\right]\right\}} |h_n| \right] \ge$$

$$\geq \sum_{\{n\in Z:\,|h_n|>1\}} 1 \;+ \sum_{k=0}^{\infty}\Big[\sum_{\{n\in Z:\,|h_n|\in\left(2^{-k-1};\,2^{-k}\right]\}} 2^{-k-1}\Big] =$$

$$= h\left(1\right) + \sum_{k=0}^{\infty}\left[2^{-k-1}\cdot\left(h\left(2^{-k-1}\right) - h\left(2^{-k}\right)\right)\right] = \sum_{k=0}^{\infty}\left[2^{-k-1}\cdot h\left(2^{-k}\right)\right]$$

implies that

$$\lim_{k\to\infty} 2^{-k}\cdot h\left(2^{-k}\right) = 0.$$

Hence, taking into account that the function $h(\lambda)$ is decreasing, we obtain (2.2).

It follows from (2.1) that, if $Hb\in l_1$, then

$$\left(Hb\right)\left(\lambda\right) = o\left(1/\lambda\right), \quad \lambda\to 0+,$$

and, therefore, by Theorem 1, we obtain that the equation (2.2) holds. The proof of Theorem 2 is complete. □

Theorem 3. *If asequence $b\in l_1$ satisfies the conditions*
(i) $\sum\limits_{n\in Z} b_n = 0$;
(ii) $\sum\limits_{m\in Z} |b_m|\ln\left(e+|m|\right) < \infty$, *then $Hb\in l_1$ and the following inequality holds*:

$$\left\|Hb\right\|_{l_1} \leq 6\sum_{m\in Z}|b_m|\ln\left(e+|m|\right). \tag{2.3}$$

P r o o f. It follows from the definition of the discrete Hilbert transform that

$$\left|\left(Hb\right)_0\right| = \Big|\sum_{m\neq 0}\frac{b_m}{m}\Big| \leq \left\|b\right\|_{l_1}. \tag{2.4}$$

From condition (i) for $n\neq 0$, we obtain that

$$\left|\left(Hb\right)_n\right| = \Big|\sum_{m\neq n}\frac{b_m}{n-m}\Big| = \Big|\sum_{m\neq n}\frac{b_m}{n-m} - \sum_{m\neq n}\frac{b_m}{n} - \frac{b_n}{n}\Big| \leq \left|\frac{b_n}{n}\right| + \sum_{m\neq n}\frac{|m|\,|b_m|}{|n|\,|n-m|}. \tag{2.5}$$

It follows from inequalities (2.4) and (2.5) that

$$\left\|Hb\right\|_{l_1} = \sum_{n\in Z}\left|\left(Hb\right)_n\right| \leq 2\left\|b\right\|_{l_1} + \sum_{n\neq 0}\Big[\sum_{m\neq n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big] =$$

$$= 2\left\|b\right\|_{l_1} + \sum_{n>0}\Big[\sum_{m>n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big] + \sum_{n>0}\Big[\sum_{m<n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big] +$$

$$+\sum_{n<0}\Big[\sum_{m>n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big] + \sum_{n<0}\Big[\sum_{m<n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big] =$$

$$= 2\left\|b\right\|_{l_1} + J_1 + J_2 + J_3 + J_4. \tag{2.6}$$

Let us estimate the summands J_k, $k=1,2,3,4$. From condition (ii) and f equalities of the form

$$\sum_{n<0}\left(\frac{1}{n-m} - \frac{1}{n}\right) = \left(\frac{1}{-1-m} + 1\right) + \left(\frac{1}{-2-m} + \frac{1}{2}\right) + \ldots + \left(\frac{1}{-m-m} + \frac{1}{m}\right) +$$

$$+\Big(\frac{1}{-m-1-m}+\frac{1}{m+1}\Big)+\Big(\frac{1}{-m-2-m}+\frac{1}{m+2}\Big)+...=1+\frac{1}{2}+...+\frac{1}{m},$$

for $m>0$, and

$$\sum_{n>0}\Big(\frac{1}{n}-\frac{1}{n-m}\Big)=\Big(1-\frac{1}{1+|m|}\Big)+\Big(\frac{1}{2}-\frac{1}{2+|m|}\Big)+...+\Big(\frac{1}{|m|}-\frac{1}{|m|+|m|}\Big)+$$

$$+\Big(\frac{1}{|m|+1}-\frac{1}{|m|+1+|m|}\Big)+\Big(\frac{1}{|m|+2}-\frac{1}{|m|+2+|m|}\Big)+...=1+\frac{1}{2}+...+\frac{1}{|m|},$$

for $m<0$, we obtain that

$$J_1=\sum_{n>0}\Big[\sum_{m>n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big]=\sum_{m>1}\Big[\sum_{0<n<m}\frac{m\,|b_m|}{n\,(m-n)}\Big]=$$

$$=\sum_{m>1}|b_m|\cdot\Big[\sum_{0<n<m}\Big(\frac{1}{m-n}+\frac{1}{n}\Big)\Big]=2\sum_{m>1}|b_m|\cdot\Big[1+\frac{1}{2}+...+\frac{1}{m-1}\Big]\le\sum_{m>1}|b_m|\cdot\ln m,$$

$$J_2=\sum_{n<0}\Big[\sum_{m>n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big]=\sum_{m>0}\Big[\sum_{n<0}\frac{m\,|b_m|}{n\,(n-m)}\Big]+\sum_{m<0}\Big[\sum_{n<m}\frac{m\,|b_m|}{n\,(m-n)}\Big]=$$

$$=\sum_{m>0}|b_m|\cdot\Big[\sum_{n<0}\Big(\frac{1}{n-m}-\frac{1}{n}\Big)\Big]+\sum_{m<0}|b_m|\cdot\Big[\sum_{n<m}\Big(\frac{1}{m-n}+\frac{1}{n}\Big)\Big]=$$

$$=\sum_{m>0}|b_m|\cdot\Big[1+\frac{1}{2}+...+\frac{1}{m}\Big]+\sum_{m<0}|b_m|\cdot\Big[1+\frac{1}{2}+...+\frac{1}{|m|}\Big]\le\sum_{m\in Z}|b_m|\cdot\ln\left(1+|m|\right),$$

$$J_3=\sum_{n>0}\Big[\sum_{m<n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big]=\sum_{m<0}\Big[\sum_{n>0}\frac{m\,|b_m|}{n\,(m-n)}\Big]+\sum_{m>0}\Big[\sum_{n>m}\frac{m\,|b_m|}{n\,(n-m)}\Big]=$$

$$=\sum_{m<0}|b_m|\cdot\Big[\sum_{n>0}\Big(\frac{1}{n}-\frac{1}{n-m}\Big)\Big]+\sum_{m>0}|b_m|\cdot\Big[\sum_{n>m}\Big(\frac{1}{n-m}-\frac{1}{n}\Big)\Big]=$$

$$=\sum_{m<0}|b_m|\cdot\Big[1+\frac{1}{2}+...+\frac{1}{|m|}\Big]+\sum_{m>0}|b_m|\cdot\Big[1+\frac{1}{2}+...+\frac{1}{m}\Big]\le\sum_{m\in Z}|b_m|\cdot\ln\left(1+|m|\right),$$

$$J_4=\sum_{n<0}\Big[\sum_{m<n}\frac{|m|\,|b_m|}{|n|\,|n-m|}\Big]=\sum_{m<-1}\Big[\sum_{m<n<0}\frac{m\,|b_m|}{n\,(n-m)}\Big]=$$

$$=\sum_{m<-1}|b_m|\cdot\Big[\sum_{m<n<0}\Big(\frac{1}{n-m}-\frac{1}{n}\Big)\Big]=$$

$$=2\sum_{m<-1}|b_m|\cdot\Big[1+\frac{1}{2}+...+\frac{1}{|m|-1}\Big]\le 2\sum_{m<-1}|b_m|\cdot\ln|m|\,.$$

From this and (2.6), we obtain (2.3). The proof of Theorem 3 is complete. □

Theorem 4. *The following equation holds under the conditions of Theorem* 3:

$$\sum_{n\in Z}\left(Hb\right)_n=0. \tag{2.7}$$

P r o o f. By the conditions of Theorem 3,

$$(Hb)_0 = -\sum_{m\neq 0}\frac{b_m}{m}$$

and

$$(Hb)_n = \sum_{m\neq n}\frac{b_m}{n-m} = \sum_{m\neq n}\frac{b_m}{n-m} - \sum_{m\neq n}\frac{b_m}{n} - \frac{b_n}{n} = \sum_{m\neq n}\frac{mb_m}{n\,(n-m)} - \frac{b_n}{n}$$

for $n\neq 0$. Therefore, we have

$$\sum_{n\in Z}(Hb)_n = -\sum_{m\neq 0}\frac{b_m}{m} + \sum_{n\neq 0}\Big[\sum_{m\neq n}\frac{mb_m}{n\,(n-m)} - \frac{b_n}{n}\Big] = -2\sum_{m\neq 0}\frac{b_m}{m} + \sum_{n\neq 0}\Big[\sum_{m\neq n}\frac{mb_m}{n\,(n-m)}\Big] =$$

$$= -2\sum_{m\neq 0}\frac{b_m}{m} + \sum_{n>0}\Big[\sum_{m>n}\frac{mb_m}{n\,(n-m)}\Big] + \sum_{n>0}\Big[\sum_{m<n}\frac{mb_m}{n\,(n-m)}\Big] +$$

$$+\sum_{n<0}\Big[\sum_{m>n}\frac{mb_m}{n\,(n-m)}\Big] + \sum_{n<0}\Big[\sum_{m<n}\frac{mb_m}{n\,(n-m)}\Big] = -2\sum_{m\neq 0}\frac{b_m}{m} + j_1 + j_2 + j_3 + j_4. \tag{2.8}$$

It follows from condition (ii) that

$$j_1 = \sum_{n>0}\Big[\sum_{m>n}\frac{mb_m}{n\,(n-m)}\Big] = \sum_{m>1}\Big[\sum_{0<n<m}\frac{mb_m}{n\,(n-m)}\Big] =$$

$$= \sum_{m>1} b_m\cdot\Big[\sum_{0<n<m}\Big(\frac{1}{n-m} - \frac{1}{n}\Big)\Big] = -2\sum_{m>1} b_m\cdot\Big[1+\frac{1}{2}+\ldots+\frac{1}{m-1}\Big],$$

$$j_2 = \sum_{n<0}\Big[\sum_{m>n}\frac{mb_m}{n\,(n-m)}\Big] = \sum_{m>0}\Big[\sum_{n<0}\frac{mb_m}{n\,(n-m)}\Big] + \sum_{m<0}\Big[\sum_{n<m}\frac{mb_m}{n\,(n-m)}\Big] =$$

$$= \sum_{m>0} b_m\cdot\Big[\sum_{n<0}\Big(\frac{1}{n-m} - \frac{1}{n}\Big)\Big] + \sum_{m<0} b_m\cdot\Big[\sum_{n<m}\Big(\frac{1}{n-m} - \frac{1}{n}\Big)\Big] =$$

$$= \sum_{m>0} b_m\cdot\Big[1+\frac{1}{2}+\ldots+\frac{1}{m}\Big] - \sum_{m<0} b_m\cdot\Big[1+\frac{1}{2}+\ldots+\frac{1}{|m|}\Big],$$

$$j_3 = \sum_{n>0}\Big[\sum_{m<n}\frac{mb_m}{n\,(n-m)}\Big] = \sum_{m<0}\Big[\sum_{n>0}\frac{mb_m}{n\,(n-m)}\Big] + \sum_{m>0}\Big[\sum_{n>m}\frac{mb_m}{n\,(n-m)}\Big] =$$

$$= \sum_{m<0} b_m\cdot\Big[\sum_{n>0}\Big(\frac{1}{n-m} - \frac{1}{n}\Big)\Big] + \sum_{m>0} b_m\cdot\Big[\sum_{n>m}\Big(\frac{1}{n-m} - \frac{1}{n}\Big)\Big] =$$

$$= -\sum_{m<0} b_m\cdot\Big[1+\frac{1}{2}+\ldots+\frac{1}{|m|}\Big] + \sum_{m>0} b_m\cdot\Big[1+\frac{1}{2}+\ldots+\frac{1}{m}\Big],$$

$$j_4 = \sum_{n<0}\Big[\sum_{m<n}\frac{mb_m}{n\,(n-m)}\Big] = \sum_{m<-1}\Big[\sum_{m<n<0}\frac{mb_m}{n\,(n-m)}\Big] =$$

$$= \sum_{m<-1} b_m\cdot\Big[\sum_{m<n<0}\Big(\frac{1}{n-m} - \frac{1}{n}\Big)\Big] = 2\sum_{m<-1} b_m\cdot\Big[1+\frac{1}{2}+\ldots+\frac{1}{|m|-1}\Big].$$

From this and (2.8), we obtain (2.7). The proof of Theorem 4 is complete.

Acknowledgements

The authors thank the anonymous referees for careful reading of the manuscript and very useful comments. This research was supported by the Science Development Foundation under the President of the Republic of Azerbaijan (grant no. EIF/MQM/Elm-Tehsil-1-2016-1(26)-71/08/01).

REFERENCES

1. Andersen K. F. Inequalities with weights for discrete Hilbert transforms. *Canad. Math. Bul.*, 1977. Vol. 20. P. 9–16.
2. Belov Y., Mengestie T. Y., Seip K. Discrete Hilbert transforms on sparse sequences. *Proc. London Math. Soc.*, 2011. Vol. 103, No. 1. P. 73–105. DOI: 10.1112/plms/pdq053
3. Belov Y., Mengestie T. Y., Seip K. Unitary discrete Hilbert transforms. *J. Anal. Math.*, 2010. Vol. 112. P. 383–393. DOI: 10.1007/s11854-010-0035-y
4. De Carli L., Samad S. One-parameter groups and discrete Hilbert transform. *Canad. Math. Bull.*, 2016. Vol. 59. P. 497–507. arXiv: 1506.03362 [math.FA]. URL: https://arxiv.org/pdf/1506.03362.pdf
5. Gabisonija I., Meskhi A. Two weighted inequalities for a discrete Hilbert transform. *Proc. A. Razmadze Math. Inst.*, 1998. Vol. 116. P. 107–122. URL: http://rmi.tsu.ge/proceedings/volumes/ps/v116-4.ps.gz
6. Hunt R., Muckenhoupt B., Wheeden R. Weighted norm inequalities for the conjugate function and Hilbert transform. *Trans. Amer. Math. Soc.*, 1973. Vol. 176, No. 2. P. 227–251. DOI: 10.2307/1996205
7. Laeng E. Remarks on the Hilbert transform and some families of multiplier operators related to it. *Collect. Math.*, 2007. Vol. 58, No. 1. P. 25–44.
URL: https://www.raco.cat/index.php/CollectaneaMathematica/article/view/57795
8. Liflyand E. Weighted Estimates for the Discrete Hilbert Transform. In: *Methods of Fourier Analysis and Approximation Theory. Applied and Numerical Harmonic Analysis*, ed. M. Ruzhansky, S. Tikhonov. Cham: Birkhäuser, 2016. P. 59–69. DOI: 10.1007/978-3-319-27466-9_5
9. Rakotondratsimba Y. Two weight inequality for the discrete Hilbert transform. *Soochow J. Math.*, 1999. Vol. 25, No. 4. P. 353–373. URL: http://mathlab.math.scu.edu.tw/mp/pdf/S25N44.pdf
10. Riesz M. Sur les fonctions conjuguees. *Math. Z.*, 1928. Vol. 27. P. 218–244.
URL: https://eudml.org/doc/167977
11. Stepanov V. D., Tikhonov S. Yu. Two weight inequalities for the Hilbert transform of monotone functions. *Dokl. Math.*, 2011. Vol. 83, No. 2. P. 241–242. DOI: 10.1134/S1064562411020359

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 13–23

DOI: 10.15826/umj.2018.2.003

ONE-SIDED L-APPROXIMATION ON A SPHERE OF THE CHARACTERISTIC FUNCTION OF A LAYER[1]

Marina V. Deikalova[†] **and Anastasiya Yu. Torgashova**[††]

Ural Federal University,
51 Lenin aven., Ekaterinburg, Russia, 620000

[†]marina.deikalova@urfu.ru, [††]anastasiya.torgashova@mail.ru

Abstract: In the space $L(\mathbb{S}^{m-1})$ of functions integrable on the unit sphere $\mathbb{S}^{m-1}$ of the Euclidean space $\mathbb{R}^m$ of dimension $m \geq 3$, we discuss the problem of one-sided approximation to the characteristic function of a spherical layer $\mathbb{G}(J) = \{x = (x_1, x_2, \ldots, x_m) \in \mathbb{S}^{m-1} \colon x_m \in J\}$, where J is one of the intervals $(a, 1]$, (a, b), and $[-1, b)$, $-1 < a < b < 1$, by the set of algebraic polynomials of given degree n in m variables. This problem reduces to the one-dimensional problem of one-sided approximation in the space $L^{\phi}(-1, 1)$ with the ultraspherical weight $\phi(t) = (1-t^2)^{\alpha}$, $\alpha = (m-3)/2$, to the characteristic function of the interval J. This result gives a solution of the problem of one-sided approximation to the characteristic function of a spherical layer in all cases when a solution of the corresponding one-dimensional problem known. In the present paper, we use results by A.G. Babenko, M.V. Deikalova, and Sz.G. Revesz (2015) and M.V. Deikalova and A.Yu. Torgashova (2018) on the one-sided approximation to the characteristic functions of intervals.

Keywords: One-sided approximation, characteristic function, spherical layer, spherical cap, algebraic polynomials.

Introduction

Let $\mathbb{R}^m,\ m \geq 2$, be the Euclidean space with the inner product

$$xy = \sum_{i=1}^{m} x_k y_k, \quad x = (x_1, x_2, \ldots, x_m), \quad y = (y_1, y_2, \ldots, y_m),$$

and the norm $\|x\| = \sqrt{xx}$. For $r > 0$, we consider in the space $\mathbb{R}^m$ the sphere $\mathbb{S}^{m-1}(r) = \{x \in \mathbb{R}^m \colon \|x\| = r\}$ of radius r centered at the origin; we denote by $\mathbb{S}^{m-1}$ the unit sphere $(r = 1)$. For $-1 \leq a < b \leq 1$, we define the intervals

$$J = \begin{cases} (a, b), & -1 < a < b < 1; \\ (a, 1], & -1 < a < b = 1; \\ [-1, b), & a = -1 < b < 1. \end{cases} \tag{0.1}$$

By means of these intervals, we define the spherical layers

$$\mathbb{G}(J) = \{x = (x_1, x_2, \ldots, x_m) \in \mathbb{S}^{m-1} \colon x_m \in J\} \tag{0.2}$$

centered at the "north pole" $e_m = (0, 0, \ldots, 0, 1)$ of the sphere. In the case $b = 1$ and $-1 < a < 1$, the layer

$$\mathbb{C}(a) = \mathbb{G}(a, 1] = \{x = (x_1, x_2, \ldots, x_m) \in \mathbb{S}^{m-1} \colon x_m \in (a, 1]\}$$

[1]This work was supported by the Russian Foundation for Basic Research (project no. 18-01-00336) and by the Russian Academic Excellence Project (agreement no. 02.A03.21.0006 of August 27, 2013, between the Ministry of Education and Science of the Russian Federation and Ural Federal University).

is the spherical cap.

Let $\mathbb{H}$ be one of the following manifolds: either an interval from the real line or a sphere $\mathbb{S}^{k-1}(r)$ of radius $r > 0$ in the space $\mathbb{R}^k$, $2 \le k \le m$, in particular, the unit sphere $\mathbb{S}^{m-1} = \mathbb{S}^{m-1}(1)$. On each of these sets, we consider the classical Lebesgue measure (of corresponding dimension). For a measurable subset $E \subset \mathbb{H}$, denote by $|E|$ the (corresponding) measure of the set E. Let $L(E)$ be the space of functions measurable and integrable on E. For a function $f \in L(E)$, its Lebesgue integral on the set E is written as $\int_E f(x)dx$. We assume that the space $L(E)$ is equipped with the norm $\|f\| = \|f\|_{L(E)} = \int_E |f(x)|dx$.

Denote by $\mathscr{P}_{n,m}$ the set of algebraic polynomials

$$P_n(x) = \sum_{\substack{|\alpha| = \alpha_1 + \cdots + \alpha_m \le n, \\ \alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{Z}_+^m}} c_\alpha x^\alpha,$$

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_m^{\alpha_m}, \quad x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m,$$

of degree (at most) n in m variables with real coefficients c_α.

In what follows, for a couple of measurable functions f and g on the sphere $\mathbb{S}^{m-1}$, the inequality $f \le g$ means that $f(x) \le g(x)$ for almost all $x \in \mathbb{S}^{m-1}$. Given a function f defined and measurable on the sphere $\mathbb{S}^{m-1}$, we consider the sets

$$\mathscr{P}^-_{n,m}(f) = \{P_n \in \mathscr{P}_{n,m} \colon P_n \le f\}, \quad \mathscr{P}^+_{n,m}(f) = \{P_n \in \mathscr{P}_{n,m} \colon P_n \ge f\} \tag{0.3}$$

of polynomials from $\mathscr{P}_{n,m}$ whose graphs "lie" under and over the graph of the function f, respectively. In order to the sets (0.3) were not empty, we assume that f is lower bounded in the former case and upper bounded in the latter case. Consider the values of the best one-sided approximation in the space $L(\mathbb{S}^{m-1})$ to the function f from below and from above by the set $\mathscr{P}_{n,m}$:

$$e^{\mp}_{n,m}(f) = \inf\{\|f - P_n\|_{L(\mathbb{S}^{m-1})} \colon P_n \in \mathscr{P}^{\mp}_{n,m}(f)\}. \tag{0.4}$$

Polynomials implementing the infimums in these relations are called *polynomials of the best* (*integral*) *approximation to the function f from below* and *from above*, respectively, or simply *extremal polynomials*. The space $\mathscr{P}_{n,m}$ is finite-dimensional; therefore, as easily understood, extremal polynomials in (0.4) exist; i. e., we can replace inf by min in (0.4) (as well as in similar problems in what follows).

The main aim of the present paper is to study the best one-sided approximation from below in the space $L(\mathbb{S}^{m-1})$ to the characteristic function

$$\mathbf{1}_{\mathbb{G}(J)}(x) = \begin{cases} 1, & x \in \mathbb{G}(J), \\ 0, & x \notin \mathbb{G}(J), \end{cases} \tag{0.5}$$

of a spherical layer (0.2) by the subspace of polynomials $\mathscr{P}_{n,m}$; more exactly, to study the value

$$e^-_{n,m}(\mathbf{1}_{\mathbb{G}(J)}) = \inf\{\|\mathbf{1}_{\mathbb{G}(J)} - P_n\|_{L(\mathbb{S}^{m-1})} \colon P_n \in \mathscr{P}^-_{n,m}(\mathbf{1}_{\mathbb{G}(J)})\}; \tag{0.6}$$

here, in accordance with the introduced notation,

$$\mathscr{P}^-_{n,m}(\mathbf{1}_{\mathbb{G}(J)}) = \{P_n \in \mathscr{P}_{n,m} \colon P_n \le \mathbf{1}_{\mathbb{G}(J)}\}.$$

The fact that function (0.5) is zonal plays an important role in what follows. In this paper, a function $f(x)$, $x = (x_1, x_2, \ldots, x_m)$, defined on the sphere $\mathbb{S}^{m-1}$ is called zonal if it depends only on the coordinate x_m of the point $x = (x_1, x_2, \ldots, x_m) \in \mathbb{S}^{m-1}$, i. e.,

$$f(x_1, x_2, \ldots, x_m) = \zeta(x_m), \quad x = (x_1, x_2, \ldots, x_m) \in \mathbb{S}^{m-1}, \tag{0.7}$$

where ζ is a univariate function defined on the interval $[-1, 1]$. For function (0.5), the function ζ in representation (0.7) is the characteristic function of the interval (0.1).

In the present paper, we only discuss problems of one-sided approximation from below. The problems of approximation from above are not specially considered. However, in certain cases, results for approximations from above can be obtained either as consequences of results on the approximation from below or by the same scheme.

Approximation without constraints in the space $L(\mathbb{S}^{m-1})$ to the characteristic function of a spherical layer and a spherical cap by algebraic polynomials was studied by one of the authors [8–10]. Note also that, by now, there are many studies devoted to approximation theory, harmonic analysis, and extremal problems on Euclidean sphere; see, for example, monographs [7, 13] and the references therein.

1. Reduction to a one-dimensional problem

In this section, we show that, by means of averaging, problems (0.4) of one-sided approximation to a zonal function in the space $L(\mathbb{S}^{m-1})$ on the sphere reduce to problems of one-sided approximation to the corresponding univariate function in the space $L^{\phi}(-1, 1)$ of functions integrable on the interval $(-1, 1)$ with the ultraspherical weight

$$\phi(t) = (1 - t^2)^{\alpha}, \quad \alpha = \frac{m-3}{2}, \tag{1.1}$$

which is equipped with the norm

$$\|g\|_{L^{\phi}(-1,1)} = \int_{-1}^{1} |g(t)|\, \phi(t)\, dt.$$

To prove this fact, we apply an averaging operator Υ, which, to a function $f \in L(\mathbb{S}^{m-1})$, set in correspondence a function of one variable $t \in (-1, 1)$ by the formula

$$(\Upsilon f)(t) = \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-2}} f\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) d\widetilde{x}. \tag{1.2}$$

The following lemma implies that this is a bounded linear operator from the space $L(\mathbb{S}^{m-1})$ to the space $L^{\phi}(-1, 1)$. This lemma is a variant of Fubini's theorem (see, for example, [12, Ch. III, Sect. 11]); a detailed proof of this statement with the passage to polar coordinates on the sphere can be found, for example, in [8].

Lemma 1. *The following statements hold for a function f integrable on the sphere $\mathbb{S}^{m-1}$ for $m \geq 3$.*

(1) *For almost all $t \in (-1, 1)$, the function $f\left(\sqrt{1-t^2}\,\widetilde{x}, t\right)$ of variable $\widetilde{x} \in \mathbb{S}^{m-2}$ is integrable on $\mathbb{S}^{m-2}$.*

(2) *The function*

$$g(t) = \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-2}} f\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) d\widetilde{x} \tag{1.3}$$

of variable $t \in (-1, 1)$ is integrable with ultraspherical weight (1.1) on the interval $(-1, 1)$; i. e., $g \in L^{\phi}(-1, 1)$.

(3) *The following formula holds:*

$$\int_{\mathbb{S}^{m-1}} f(x)dx = |\mathbb{S}^{m-2}| \int_{-1}^{1} g(t)(1 - t^2)^{(m-3)/2}\, dt. \tag{1.4}$$

According to Lemma 1, the averaging operator Υ is well defined by formula (1.2) on the whole space $L(\mathbb{S}^{m-1})$ and is a linear operator from $L(\mathbb{S}^{m-1})$ to $L^{\phi}(-1,1)$. The following estimate holds for $f \in L(\mathbb{S}^{m-1})$ for almost all points $t \in (-1,1)$:

$$|g(t)| = |(\Upsilon f)(t)| \le \widetilde{g}(t) = \frac{1}{|\mathbb{S}^{m-2}|}\int_{\mathbb{S}^{m-2}} \left| f\left(\sqrt{1-t^2}\,\widetilde{x},\, t\right)\right| d\widetilde{x} = (\Upsilon |f|)(t).$$

Hence, using formula (1.4) for the function $|f|$, we obtain the inequality

$$|\mathbb{S}^{m-2}|\ \|g\|_{L^{\phi}(-1,1)} = |\mathbb{S}^{m-2}| \int_{-1}^{1} |g(t)|\,(1-t^2)^{(m-3)/2}\,dt \le \|f\|_{L(\mathbb{S}^{m-1})}. \tag{1.5}$$

The latter inequality turns into an equality at least in the following two cases: (1) the function f is nonnegative; (2) the function f is zonal, more exactly, depends only on the variable x_m. Indeed, if f is nonnegative, then the unique inequality in the chain of relations given above turns into an equality. In the case when the function f is zonal, we have $\Upsilon f = f$, i. e., $f(x) = g(x_m)$ and, obviously, inequality (1.5) turns into an equality.

Thus, the averaging operator (1.2) is a bounded linear operator from $L(\mathbb{S}^{m-1})$ to $L^{\phi}(-1,1)$ and the equality $|\mathbb{S}^{m-2}|\ \|\Upsilon\| = 1$ holds for its norm $\|\Upsilon\|$.

Note also that, if the function f is continuous on the sphere $\mathbb{S}^{m-1}$, then the function $g = \Upsilon f$ is continuous on the interval $\mathbb{I} = [-1,1]$ and $g(1) = f(e_m)$, $e_m = (0,0,\dots,0,1)$.

The following lemma describes the structure of function (1.3) in the case when f is a polynomial. The proof of the lemma can also be found in [8].

Lemma 2. *For $n \ge 1$ and $m \ge 3$, for any polynomial $P_n \in \mathscr{P}_{n,m}$, the function*

$$g_n(t) = (\Upsilon P_n)(t) = \frac{1}{|\mathbb{S}^{m-2}|}\int_{\mathbb{S}^{m-2}} P_n\left(\sqrt{1-t^2}\,\widetilde{x},\, t\right) d\widetilde{x} \tag{1.6}$$

is a polynomial in (one) variable $t = x_m$ of degree at most n.

Thus, we have the embedding $\Upsilon \mathscr{P}_{n,m} \subset \mathscr{P}_n$, where $\mathscr{P}_n = \mathscr{P}_{n,1}$ is the set of polynomials in one variable of degree at most n. In fact, we have the equality

$$\Upsilon \mathscr{P}_{n,m} = \mathscr{P}_n. \tag{1.7}$$

Indeed, a polynomial $p_n \in \mathscr{P}_n$ can be regarded as a polynomial in m variables; more exactly, with the polynomial p_n, we associate the polynomial $P_n(x_1, x_2, \dots, x_m) = p_n(x_m)$ in m variables. In this case, the right-hand side of (1.6) is the polynomial p_n; thus, $\Upsilon P_n = p_n$. Relation (1.7) is verified.

Lemma 3. *Let $m \ge 3$. If a function f is defined, integrable, and lower bounded on the sphere $\mathbb{S}^{m-1}$, then the embedding $\Upsilon(\mathscr{P}^{-}_{n,m}(f)) \subset \mathscr{P}^{-}_n(\Upsilon f)$ holds. If, in addition, the function f is zonal, then the following equality holds:*

$$\Upsilon(\mathscr{P}^{-}_{n,m}(f)) = \mathscr{P}^{-}_n(\Upsilon f).$$

P r o o f. For the beginning, let f be an arbitrary integrable and lower bounded (not necessarily zonal) function on the sphere, and let $P_n \in \mathscr{P}^{-}_{n,m}(f)$. Recall that the property $P_n \in \mathscr{P}^{-}_{n,m}(f)$ means that the set $\Omega = \Omega(f, P_n)$ of points $x \in \mathbb{S}^{m-1}$ at which the inequality $P_n(x) \le f(x)$ holds is of full measure (i. e., the difference $\mathbb{S}^{m-1} \setminus \Omega$ is a null-measure set) in $\mathbb{S}^{m-1}$.

We have to compare the averaging of the polynomial P_n

$$(\Upsilon P_n)(t) = \frac{1}{|\mathbb{S}^{m-2}|}\int_{\mathbb{S}^{m-2}} P_n\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) d\widetilde{x} \tag{1.8}$$

and the averaging of the function f

$$(\Upsilon f)(t) = \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-2}} f\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) d\widetilde{x} \tag{1.9}$$

for (almost all) $t \in (-1,1)$.

Since a polynomial P_n is a continuous function, averaging (1.8) exists for all points $t \in [-1,1]$; moreover, according to Lemma 2, $(\Upsilon P_n)(t)$ is a polynomial in one variable. Averaging (1.9) exists, in general, not for all $t \in (-1,1)$. According to Lemma 1, integrals (1.9) exist for almost all $t \in (-1,1)$, i. e., everywhere on $(-1,1)$ except for a null-measure set $I_0' \subset (-1,1)$.

Let us study the set of pairs $(\widetilde{x}, t)$ of points at which the following inequality holds:

$$P_n\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) \le f\left(\sqrt{1-t^2}\,\widetilde{x}, t\right).$$

Let us ascertain that this inequality holds for almost all $t \in (-1,1)$ (i. e., everywhere on $(-1,1)$ except for a null-measure set I_0'') for almost all $\widetilde{x} \in \mathbb{S}^{m-2}$ (i. e., everywhere on $\mathbb{S}^{m-2}$ except for a null-measure set $S_0(t)$). In other words, the set

$$\omega(t) = \left\{\widetilde{x} \in \mathbb{S}^{m-2} \colon \left(\sqrt{1-t^2}\,\widetilde{x}, t\right) \in \Omega\right\} \tag{1.10}$$

is of full measure in $\mathbb{S}^{m-2}$ for almost all $t \in (-1,1)$. This fact can be proved, for example, on the following way.

Let us apply Lemma 1 to the characteristic function $\mathbf{1}_\Omega$ of the set Ω. By the first statement of the lemma, the set $\omega(t)$ is measurable on $\mathbb{S}^{m-2}$ for almost all $t \in (-1,1)$ (i. e., everywhere on $(-1,1)$ except for a null-measure set I_0''). Formula (1.4) for the function $\mathbf{1}_\Omega$ takes the form

$$|\mathbb{S}^{m-1}| = \int_{-1}^{1} |\omega(t)|_{m-2} (1-t^2)^{(m-3)/2}\, dt. \tag{1.11}$$

For the measure of set (1.10), we have the estimate

$$|\omega(t)|_{m-2} \le |\mathbb{S}^{m-2}|. \tag{1.12}$$

Therefore, (1.11) implies that

$$|\mathbb{S}^{m-1}| = \int_{-1}^{1} |\omega(t)|_{m-2} (1-t^2)^{(m-3)/2}\, dt \le |\mathbb{S}^{m-2}| \int_{-1}^{1} (1-t^2)^{(m-3)/2}\, dt.$$

The latter value is $|\mathbb{S}^{m-1}|$. Hence, inequality (1.12) turns into an equality on the set $(-1,1) \setminus I_0''$. Thus, set (1.10) is measurable for almost all $t \in (-1,1)$ and

$$|\omega(t)|_{m-2} = |\mathbb{S}^{m-2}|. \tag{1.13}$$

The set $I_0 = I_0' \bigcup I_0'' \subset (-1,1)$ is a null-measure set; (1.9) holds for points $t \notin I_0$, the set (1.10) is measurable, and (1.13) hods. For points $t \in (-1,1) \setminus I_0$, the inequality $(\Upsilon P_n)(t) \le (\Upsilon f)(t)$ holds.

Consequently, the embedding $\Upsilon(\mathscr{P}^-_{n,m}(f)) \subset \mathscr{P}^-_n(\Upsilon f)$ holds. Most likely, the inverse embedding does not hold in the general case.

Assume that a function f is zonal and, moreover, can be represented in the form (0.7). A polynomial $p_n \in \mathscr{P}^-_n(g)$ regarded as a zonal polynomial, obviously, belongs to the set $\mathscr{P}^-_{n,m}(f)$ and $\Upsilon p_n = p_n$. Therefore, in this case, the inverse embedding $\mathscr{P}^-_n(g) \subset \Upsilon(\mathscr{P}^-_{n,m}(f))$ hods. Lemma 3 is proved. □

For a function $g \in L^\phi(-1,1)$ lower bounded on $(-1,1)$, consider the value

$$E^-_{n,\phi}(g) = \inf\{\|g-p\|_{L^\phi(-1,1)} \colon p \in \mathscr{P}^-_n(f)\} \tag{1.14}$$

of its best approximation from below in the space $L^\phi(-1,1)$ by the set $\mathscr{P}_n$.

Theorem 1. *Let $m \geq 3$ and $n \geq 0$. For an arbitrary function $f \in L(\mathbb{S}^{m-1})$ lower bounded on $\mathbb{S}^{m-1}$, the following inequality holds:*

$$e^-_{n,m}(f) \geq |\mathbb{S}^{m-2}|\, E^-_{n,\phi}(\Upsilon f). \tag{1.15}$$

For a zonal function $f \in L(\mathbb{S}^{m-1})$, the following equality holds:

$$e^-_{n,m}(f) = |\mathbb{S}^{m-2}|\, E^-_{n,\phi}(\Upsilon f) \tag{1.16}$$

and, if a polynomial p^*_n in one variable is extremal in problem (1.14) for the function $g = \Upsilon f$, (i. e., the infimum in (1.14) is attained at this polynomial), then the zonal polynomial $P^*_n(x) = p^*_n(x_m)$, $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, is extremal in problem (0.6).

P r o o f. Let $P_n \in \mathscr{P}^-_{n,m}(f)$, and let g_n be the function (of one variable) constructed by formula (1.3):

$$g_n(t) = \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-2}} P_n\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) d\widetilde{x}.$$

As shown above, $g_n \in \mathscr{P}^-_n(g)$, where $g = \Upsilon f$. We have

$$g(t) - g_n(t) = \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-2}} \left(g(t) - P_n\left(\sqrt{1-t^2}\,\widetilde{x}, t\right)\right) d\widetilde{x}.$$

Hence,

$$|\mathbb{S}^{m-2}| \times \|g - g_n\|_{L^\phi(-1,1)} = \|f - P_n\|_{L(\mathbb{S}^{m-1})}.$$

Indeed, the following chain of relations holds:

$$\|g - g_n\|_{L^\phi(-1,1)} = \int_{-1}^{1} |g(t) - g_n(t)|(1-t^2)^{(m-3)/2}\, dt =$$

$$= \int_{-1}^{1} (g(t) - g_n(t))(1-t^2)^{(m-3)/2}\, dt =$$

$$= \int_{-1}^{1} (1-t^2)^{(m-3)/2} \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-2}} \left(f\left(\sqrt{1-t^2}\,\widetilde{x}, t\right) - P_n\left(\sqrt{1-t^2}\,\widetilde{x}, t\right)\right) d\widetilde{x}\, dt =$$

$$= \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-1}} (f(x) - P_n(x))\, dx = \frac{1}{|\mathbb{S}^{m-2}|} \int_{\mathbb{S}^{m-1}} |f(x) - P_n(x)|\, dx =$$

$$= \frac{1}{|\mathbb{S}^{m-2}|} \|f - P_n\|_{L(\mathbb{S}^{m-1})}.$$

This, by Lemma 3, implies relations (1.15) and (1.16). Due to equality (1.16), at the polynomial $P^*_n(x) = p^*_n(x_m)$, $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, the infimum in (0.6) is attained; thus, this polynomial is extremal. Theorem 1 is proved. □

Function (0.5) is zonal; more exactly,

$$\mathbf{1}_{\mathbb{G}(J)}(x) = \mathbf{1}_J(x_m), \quad x = (x_1, x_2, \ldots, x_m) \in \mathbb{S}^{(m-1)},$$

where $\mathbf{1}_J$ is the characteristic function of the interval (0.1):

$$\mathbf{1}_J(t) = \begin{cases} 1, & t \in J, \\ 0, & t \in [-1,1] \setminus J. \end{cases}$$

Consider the best approximation from below

$$E^-_{n,\phi}(\mathbf{1}_J) = \inf\{\|\mathbf{1}_J - p_n\|_{L^\phi(-1,1)} : p_n \in \mathscr{P}^-_n(\mathbf{1}_J)\} \tag{1.17}$$

to the step function $\mathbf{1}_J$ in the space $L^\phi(-1,1)$ by the set $\mathscr{P}^-_n(\mathbf{1}_J) = \mathscr{P}^-_{n,1}(\mathbf{1}_J)$ of algebraic polynomials (in one variable) whose graphs lie under the graph of the function $\mathbf{1}_J$. As a particular case of Theorem 1, the following statement is valid.

Theorem 2. *For any $m \geq 3$ and $n \geq 0$, the following formula holds for intervals* (0.1):

$$e^-_{n,m}(\mathbf{1}_{\mathbb{G}(J)}) = |\mathbb{S}^{m-2}|\, E^-_{n,\phi}(\mathbf{1}_J)$$

and, if a polynomial p^*_n in one variable is extremal in problem (1.17)*, then the zonal polynomial $P^*_n(x) = p^*_n(x_m),\ x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, is extremal in problem* (0.6).

2. One-sided approximation to the characteristic function of an interval

Let υ be a measurable integrable nonnegative function different from zero almost everywhere on $(-1,1)$; we will call such a function a weight (on $(-1,1)$). Denote by $L^\upsilon(-1,1)$ the space of real-valued functions f integrable with weight υ on $(-1,1)$ equipped with the norm

$$\|f\| = \|f\|_{L^\upsilon(-1,1)} = \int_{-1}^{1} |f(t)|\upsilon(t)\,dt.$$

In this section, we give results from [1, 5, 11] on the one-sided approximation in the space $L^\upsilon(-1,1)$ to the characteristic function of an interval by algebraic polynomials. The results from Section 1 and from this sections make it possible to find the best one-sided approximation in the space $L(\mathbb{S}^{m-1})$ to the characteristic function of a spherical layer (in particular, a spherical cap) by algebraic polynomials in certain situations.

For nonnegative integer n, we denote by $\mathscr{P}_n$ the set of algebraic polynomials $p(t) = \sum_{k=0}^{n} a_k t^k$ in one real variable of degree at most n with real coefficients.

In contrast to Section 1, in this section, for a couple of measurable functions f and g on the interval $[-1,1]$, the inequality $f \leq g$ means that $f(t) \leq g(t)$ for all $t \in [-1,1]$. For a function f defined, bounded, and measurable on the interval $[-1,1]$, we consider the sets

$$\mathscr{P}^-_n(f) = \{p \in \mathscr{P}_n : p \leq f\}, \quad \mathscr{P}^+_n(f) = \{p \in \mathscr{P}_n : p \geq f\}$$

of polynomials from $\mathscr{P}_n$ whose graphs lie under and over over the graph of the function f, respectively. The function f is assumed lower bounded in the former case and upper bounded in the latter case. We are interested in the values

$$E^\mp_{n,\upsilon}(f) = \inf\{\|f - p\|_{L^\upsilon(-1,1)} : p \in \mathscr{P}^\mp_n(f)\} \tag{2.1}$$

of the best approximation in the space $L^\upsilon(-1,1)$ to the function f from below and from above by the set $\mathscr{P}_n$ as well as in extremal polynomials at which the infimums in (2.1) are attained.

An important tool for studying the one-sided approximation to functions by polynomials are quadrature formulas

$$\int_{-1}^{1} \upsilon(t)p(t)\,dt = \sum_{k=1}^{M} \lambda_k p(t_k), \qquad p \in \mathscr{P}_n, \tag{2.2}$$

exact on the set of polynomials $\mathscr{P}_n$ with nodes $-1 \leq t_1 < t_2 < \cdots < t_M \leq 1$ and positive weights: $\lambda_k > 0,\ 1 \leq k \leq M$; such quadrature formulas are called positive. The largest degree n of polynomials for which formula (2.2) holds is called the degree of precision of this formula. Depending on a situation, some of nodes of formula (2.2) can be fixed while the remaining are assumed free; more exactly, they are chosen so that the formula have the highest degree of precision (see, for example, [15, Ch. 7, Sect. 1]. The following statement is due to Bojanic and DeVore [4, the proof of Theorem 2] (see also [14, Theorem 1.7.5]).

Theorem A. *Assume that quadrature formula (2.2) holds on the set $\mathscr{P}_n$. Then, the following estimates are valid for a measurable bounded function f:*

$$E^-_{n,\upsilon}(f) \geq \int_{-1}^{1} \upsilon(t)f(t)\,dt - \sum_{k=1}^{M} \lambda_k f(t_k), \quad E^+_{n,\upsilon}(f) \geq \sum_{k=1}^{M} \lambda_k f(t_k) - \int_{-1}^{1} \upsilon(t)f(t)\,dt. \tag{2.3}$$

If an inequality in (2.3) turns into an equality, then quadrature formula (2.2) is said to be extremal in the corresponding problem (2.1).

Consider the problem of one-sided approximation to the characteristic function

$$\mathbf{1}_J(t) = \begin{cases} 1, & t \in J, \\ 0, & t \in [-1,1] \setminus J, \end{cases}$$

of interval (0.1) in the space $L^\upsilon(-1,1)$ by algebraic polynomials of given degree $n \geq 0$. The problem consists in finding the values

$$E^\mp_n(\mathbf{1}_J) = E^\mp_{n,\upsilon}(\mathbf{1}_J) = \inf\{\|\mathbf{1}_J - p_n\|_{L^\upsilon(-1,1)} \colon p_n \in \mathscr{P}^\mp_n(\mathbf{1}_J)\}. \tag{2.4}$$

Problems of one-sided weighted integral approximation to the characteristic function of an interval and to similar functions by algebraic or trigonometric polynomials arise in various areas of mathematics and have a rich history (see [1, 2, 5, 6, 11, 14, 16, 17] and the references therein). Let us outline only several exact results on problem (2.4) closely related to the present paper; for a more complete presentation of the topic see [1, 11]. Problem (2.4) of one-sided integral approximation to the characteristic function of an arbitrary half-open interval $(a,1] \subset (-1,1]$ by algebraic polynomials on $[-1,1]$ with the unit weight was solved, and the whole class of extremal polynomials was described in [5]. This problem in the space $L^\upsilon(-1,1)$ with an arbitrary weight is solved in [1]. Let us describe the main result of [1] in the form convenient for us.

In the study of problems (2.4) of one-sided approximation to the characteristic function of an interval by polynomials, M-point quadrature formulas are used, the set $\mathfrak{u}$ of fixed nodes of which either contains no fixed nodes or contains one, two, or three fixed nodes of a specific form:

$$\varnothing, \quad \{-1\}, \quad \{1\}, \quad \{-1,1\}, \tag{2.5}$$

$$\{\theta\}, \quad \{-1,\theta\}, \quad \{\theta,1\}, \quad \{-1,\theta,1\}, \quad \text{where} \quad \theta \in (-1,1), \tag{2.6}$$

and other $M - |\mathfrak{u}|$ nodes are chosen so that the formula have the highest degree of precision; here, $|\mathfrak{u}|$ is the cardinality, i. e., the number of points of the set $\mathfrak{u}$. It is known (see, for example, [15, Ch. 7, Sect. 1]) that the degree of precision of such formula is $n = 2M - 1 - |\mathfrak{u}|$. Formulas (2.2) take the form

$$\int_{-1}^{1} \upsilon(t)p(t)dt = \sum_{k=1}^{M} \lambda_k p(t_k), \qquad p \in \mathscr{P}_{2M-1-|\mathfrak{u}|}; \tag{2.7}$$

in what follows, we sometimes will use more accurate (in comparison with (2.2)) notation for nodes $\{t_k = t_k^{\mathfrak{u}} = t_k(\mathfrak{u},\upsilon,M)\}_{k=1}^{M}$ and weights (coefficients) $\{\lambda_k = \lambda_k^{\mathfrak{u}} = \lambda_k(\mathfrak{u},\upsilon,M)\}_{k=1}^{M}$ in formula (2.7).

In the case of the empty set $\mathfrak{u}$ (there are no fixed nodes), formula (2.7) is the classical Gauss quadrature formula (see [15, Ch. 7, Sect. 1])). In the cases $\mathfrak{u} = \{-1\}$ and $\mathfrak{u} = \{1\}$, formula (2.7) is the left and right Radau quadrature formula, respectively; in the case $\mathfrak{u} = \{-1, 1\}$, (2.7) is the Lobatto quadrature formula. It is known (see the references in [1, 3, 5]) that formula (2.7) is positive in all these cases.

For each of the sets $\mathfrak{u}$ of fixed nodes from (2.6), the set $\Theta_M^{\mathfrak{u}}$ of values of the parameter $\theta \in (-1, 1)$ for which quadrature formula (2.7) has positive weights is described in [3, 5]. Such formulas are called *quasi Gauss, quasi (left and right) Radau, and quasi Lobatto positive quadrature formulas*. In what follows, we consider formula (2.7) with fixed nodes (2.6) only for $\theta \in \Theta_M^{\mathfrak{u}}$. Thus, a quadrature formula of the form (2.7) with fixed nodes (2.5) and with fixed nodes (2.6) is positive. The degree of precision of formula (2.7) with fixed nodes (2.5) and with fixed nodes (2.6) is $N = 2M - 1 - |\mathfrak{u}|$.

The best approximation from below

$$E_{n,\upsilon}^{-}(\mathbf{1}_{(a,1]}) = \min\{\|\mathbf{1}_{(a,1]} - p_n\|_{L^{\upsilon}(-1,1)} : p_n \in \mathscr{P}_n^{-}(\mathbf{1}_{(a,1]})\} \tag{2.8}$$

and an extremal polynomial $p_n^a = p_{n,a}^{\upsilon}$ at which the minimum in (2.8) is attained were found for all values $a \in (-1, 1)$ and $n \geq 1$ in the case of the unit weight $\upsilon \equiv 1$ in [5] and in the case of an arbitrary weight υ in [1]. Results of several statements from [1, Sect. 3] containing the solution of problem (2.8) in the form convenient for us are gathered in the following theorem.

Theorem B [1, Sect. 3]. *For $M \in \mathbb{N}$, $M \geq 3$, the following statements hold.*

(1) *If the number $a \in (-1, 1)$ coincides with one of the nodes of an M-point positive quadrature formula (2.7) different from the maximum node, i. e., $a = t_{\nu}^{\mathfrak{u}}$, $1 \leq \nu \leq M - 1$, then*

$$E_{n,\upsilon}^{-}(\mathbf{1}_{(a,1]}) = \int_{(a,1]} \upsilon(t)\, dt - \sum_{k=\nu+1}^{M} \lambda_k^{\mathfrak{u}}$$

for $n = 2M - 2 - |\mathfrak{u}|$ and $n = 2M - 1 - |\mathfrak{u}|$ in the case of fixed nodes (2.5) and for $n = 2M - 1 - |\mathfrak{u}|$ in the case of fixed nodes (2.6). Moreover, the corresponding quadrature formula is extremal, and the polynomial of the best approximation from below is the polynomial $p_n^a \in \mathscr{P}_n^{-}(\mathbf{1}_{(a,1]})$ of degree $n = 2M - 2 - |\mathfrak{u}|$ for $\mathfrak{u}$ from (2.5) and of degree $n = 2M - 1 - |\mathfrak{u}|$ for $\mathfrak{u}$ from (2.6) that interpolates the function $\mathbf{1}_{(a,1]}$ at the nodes of the quadrature formula.

(2) *If the maximum node $t_M^{\mathfrak{u}}$ of formula (2.7) is less than 1, then*

$$E_{n,\upsilon}^{-}(\mathbf{1}_{(a,1]}) = \int_{(a,1]} \upsilon(t)\, dt$$

for $t_M^{\mathfrak{u}} \leq a < 1$ for all $0 \leq n \leq 2M - 1 - |\mathfrak{u}|$, and $p^* \equiv 0$ is the polynomial of the best approximation from below.

Remark. A statement similar to Theorem B is also valid for the problem $E_{n,\upsilon}^{-}(\mathbf{1}_{[-1,b)})$ of the best approximation from below to the characteristic function $\mathbf{1}_{[-1,b)}$ of an interval $[-1, b)$. In what follows, we denote by q_n^b the extremal polynomial in this problem.

Theorem B and its analog for an interval $[-1, b)$ make it possible to obtain a solution of the problem

$$E_{n,\upsilon}^{-}(\mathbf{1}_{(a,b)}) = \inf\{\|\mathbf{1}_{(a,b)} - p_n\|_{L^{\upsilon}(-1,1)} : p_n \in \mathscr{P}_n^{-}(\mathbf{1}_{(a,b)})\}$$

for intervals (a, b) whose end-points are nodes of quadrature formula (2.7). The following statement was obtained in the authors' paper [11].

Theorem C [11, Theorem 2]. *If numbers a and b, $-1 < a < b < 1$, are nodes of an M-point positive quadrature formula (2.7), more exactly,*

$$a = t_{k(a)}^{\mathfrak{u}}, \qquad b = t_{k(b)}^{\mathfrak{u}}, \qquad k(a) < k(b),$$

then

$$E^{-}_{n,\upsilon}(\mathbf{1}_{(a,b)}) = \int_a^b \upsilon(t)\,dt - \sum_{k(a)<k<k(b)} \lambda_k^{\mathfrak{u}}$$

for $n = 2M-2-|\mathfrak{u}|$ *and* $n = 2M-1-|\mathfrak{u}|$ *in the case of fixed nodes* (2.5) *and for* $n = 2M-1-|\mathfrak{u}|$ *in the case of fixed nodes* (2.6)*. Moreover, the corresponding quadrature formula is extremal, and the polynomial of the best approximation from below is the polynomial* $\varrho_n^{ab} = p_n^a + q_n^b - 1$ *of degree* $n = 2M-2-|\mathfrak{u}|$ *in the case of fixed nodes* (2.5) *and of degree* $n = 2M-1-|\mathfrak{u}|$ *in the case of fixed nodes* (2.6)*.*

3. One-sided approximation to the characteristic function of a spherical cap and a spherical layer

Theorem 2 and the results of [1, 11] presented in Section 2 give a solution of problem (0.6) of the best approximation from below in the space $L(\mathbb{S}^{m-1})$ to the characteristic function $\mathbf{1}_{\mathbb{G}(J)}$ of spherical layer (0.2) by the set $\mathscr{P}_{n,m}$ of algebraic polynomials of degree n in m variables for $m \geq 3$ in the following cases.

(1) *Approximation from below in* $L(\mathbb{S}^{m-1})$ *to the characteristic function* $\mathbf{1}_{\mathbb{C}(a)}$ *of a spherical cap* $\mathbb{C}(a) = \mathbb{G}(a,1]$ *for* $a \in (-1,1)$ *which is a node of quadrature formula* (2.7)*.* A solution is provided by Theorem 2 and the one-dimensional results from [1] given in Theorem B for the ultraspherical weight

$$\upsilon(t) = \phi(t) = (1-t^2)^{\alpha}, \quad \alpha = (m-3)/2. \tag{3.1}$$

(2) *Approximation from below in* $L(\mathbb{S}^{m-1})$ *to the characteristic function* $\mathbf{1}_{\mathbb{G}(a,b)}$ *of a spherical layer* $\mathbb{G}(a,b)$*, where* $-1 < a < b < 1$ *are nodes of positive quadrature formula* (2.7)*.* To obtain a solution of this problem, we apply the results of [11] described in Theorem C for the ultraspherical weight (3.1) and Theorem 2.

4. Conclusion

Theorem 2 gives a solution of problem (0.6) in all cases when a solution of the corresponding one-dimensional problem is known. For example, Theorem 5 from [11] for the weight υ chosen from the condition $\upsilon(t^2)|t| = \phi(t)$, $t \in (-1,1)$, i. e., for the Jacobi weight $\upsilon(t) = t^{-1/2}(1-t)^{(m-3)/2}$, $t \in (0,1)$, together with Theorem 2 give the best approximation from below in $L(\mathbb{S}^{m-1})$ to the characteristic function $\mathbf{1}_{\mathbb{G}(-h,h)}$ of a spherical layer $\mathbb{G}(-h,h)$ symmetric with respect to "equator" $x_m = 0$ of the sphere for $h \in (0,1)$.

The results of Section 1, including Theorem 2, can be naturally transferred to the problem of one-sided approximation in the space $L(\mathbb{S}^{m-1})$ from above. These results and the corresponding one-dimensional results on the approximation from above to the characteristic functions of intervals from [1, 11], make it possible to obtain a write solution of the problem $e^{+}_{n,m}(\mathbf{1}_{\mathbb{G}(J)})$ on the best approximation from above in the space $L(\mathbb{S}^{m-1})$ to the characteristic function $\mathbf{1}_{\mathbb{G}(J)}$ of spherical layer (0.2) by the set $\mathscr{P}_{n,m}$ of algebraic polynomials of degree n in m variables, at least, in the two cases described in Section 3.

Acknowledgements

The authors are grateful to Professor V.V. Arestov for the attention to their study and useful discussion of the results.

REFERENCES

1. Babenko A. G., Deikalova M. V., Revesz Sz. G. Weighted one-sided integral approximations to characteristic functions of intervals by polynomials on a closed interval. *Proc. Steklov Inst. Math.*, 2017. Vol. 297, Suppl. 1. P. S11–S18. DOI: 10.1134/S0081543817050029
2. Babenko A. G., Kryakin Yu. V., Yudin V. A., One-sided approximation in L of the characteristic function of an interval by trigonometric polynomials. *Proc. Steklov Inst. Math.*, 2013. Vol. 280, Suppl. 1. P. S39–S52. DOI: 10.1134/S0081543813020041
3. Beckermann B., Bustamante J., Martinez-Cruz R., Quesada J. M. Gaussian, Lobatto and Radau positive quadrature rules with a prescribed abscissa. *Calcolo*, 2014. Vol. 51, No. 2. P. 319–328. DOI: 10.1007/s10092-013-0087-3
4. Bojanic R., DeVore R. On polynomials of best one-sided approximation. *Enseign. Math.*, 1966. Vol. 12. P. 139–164.
5. Bustamante J., Martínez-Cruz R., Quesada J. M. Quasi orthogonal Jacobi polynomials and best one-sided L_1 approximation to step functions. *J. Approx. Theory*, 2015. Vol. 198. P. 10–23. DOI: 10.1016/j.jat.2015.05.001
6. Bustamante J., Quesada J.M., Martínez-Cruz R. Best one-sided L_1 approximation to the Heaviside and sign functions. *J. Approx. Theory*, 2012. Vol. 164, No. 6. P. 791–802. DOI: 10.1016/j.jat.2012.02.006
7. Dai F., Xu Y. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. New York: Springer Science & Business Media, 2013. 440 p. DOI: 10.1007/978-1-4614-6660-4
8. Deikalova M. V. The Taikov functional in the space of algebraic polynomials on the multidimensional Euclidean sphere. *Math Notes*, 2008. Vol. 84, No. 3–4. P. 498–514. DOI: 10.1134/S0001434608090228
9. Deikalova M. V. Integral approximation of the characteristic function of a spherical cap by algebraic polynomials. *Proc. Steklov Inst. Math.*, 2011. Vol. 273, Suppl. 1. P. S74–S85. DOI: 10.1134/S0081543811050087
10. Deikalova M. V. Several extremal approximation problems for the characteristic function of a spherical layer. *Proc. Steklov Inst. Math.*, 2012. Vol. 277, Suppl. 1. P. S79–S92. DOI: 10.1134/S0081543812050094
11. Deikalova M. V., Torgashova A. Yu. One-sided integral approximation of the characteristic function of an interval by algebraic polynomials. *Trudy Inst. Mat. i Mekh. UrO RAN* [Proc. of Krasovskii Institute of Mathematics and Mechanics of the UB RAS], 2018. Vol. 24, No. 4. P. 110–125. (In Russian) DOI: 10.21538/0134-4889-2018-24-4-110-125
12. Dunford N., Schwartz J. T. *Linear Operators. Part I: General Theory*. New York: Wiley-Interscience, 1988. 872 p.
13. Gorbachev D. V. *Izbrannye zadachi teorii funkcij i teorii priblizhenij i ih prilozheniya* [Selected Problems in Functional Analysis and Approximation Theory and Their Applications]. Tula: TulGU, 2004. 152 p. (In Russian)
14. Korneichuk N. P., Ligun A. A., Doronin V. G. *Approksimaciya s ogranicheniyami* [Approximation with Constraints]. Kiev: Naukova Dumka, 1982. 254 p. (In Russian)
15. Krylov V. I. *Approximate Calculation of Integrals*. Mineola, New York: Dover Publ. 2006. 368 p.
16. Li X.-J., Vaaler J. D. Some trigonometric extremal functions and the Erdös–Turán type inequalities. *Indiana Univ. Math. J.*, 1999. Vol. 48, No. 1. P. 183–236. DOI: 10.1512/iumj.1999.48.1508
17. Motornyi V. P., Motornaya O. V., Nitiema P. K. One-sided approximation of a step by algebraic polynomials in the mean. *Ukrainian Math. J.*, 2010. Vol. 62, No. 3. P. 467–482. DOI: 10.1007/s11253-010-0366-y

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 24–32
DOI: 10.15826/umj.2018.2.004

ORDER EQUALITIES IN DIFFERENT METRICS FOR MODULI OF SMOOTHNESS OF VARIOUS ORDERS

Niyazi A. Il'yasov

Baku State University, Baku, AZ 1148, Azerbaijan

niyazi.ilyasov@gmail.com

Abstract: In this paper, we obtain order equalities for the kth order $L_q(T)$-moduli of smoothness $\omega_k(f;\delta)_q$ in terms of expressions that contain the lth order $L_p(T)$-moduli of smoothness $\omega_l(f;\delta)_p$ on the class of periodic functions $f \in L_p(T)$ with monotonically decreasing Fourier coefficients, where $1 < p < q < \infty$, $k, l \in \mathbb{N}$, and $T = (-\pi, \pi]$.

Keywords: Inequalities of different metrics for moduli of smoothness, Order equality, Trigonometric Fourier series with monotone coefficients.

Let $L_p(\mathbb{T})$, $1 \le p < \infty$, be the space of all measurable 2π-periodic functions with finite $L_p(\mathbb{T})$-norm

$$\|f\|_p = \Big(\pi^{-1} \int_{\mathbb{T}} |f(x)|^p \, dx\Big)^{1/p},$$

where $\mathbb{T} = (-\pi, \pi]$; let $E_n(f)_p$ be the best approximation of a function f in the metric $L_p(\mathbb{T})$ by trigonometric polynomials of order at most n, $n \in \mathbb{Z}_+$; and let $\omega_l(f;\delta)_p$, where $l \in \mathbb{N}$ and $\delta \in [0, +\infty)$, be the lth order modulus of smoothness of a function $f \in L_p(\mathbb{T})$:

$$\omega_l(f;\delta)_p = \sup\big\{\|\Delta_h^l f(\cdot)\|_p \colon h \in \mathbb{R},\ |h| \le \delta\big\},$$

where

$$\Delta_h^l f(x) = \sum_{\nu=0}^{l} (-1)^{l-\nu} \binom{l}{\nu} f(x + \nu h), \quad \binom{l}{\nu} = \frac{l!}{\nu!(l-\nu)!}, \quad \nu = \overline{0, l}.$$

The following statement contains known upper estimates for $\omega_k(f;\delta)_q$ in terms of $\omega_l(f;\delta)_p$, where $f \in L_p(\mathbb{T})$, $p < q$, and $l, k \in \mathbb{N}$ (see, for example, [3, Theorem 1]; the background and the corresponding references can also be found in [3]).

Theorem A. *Let $1 \le p < q < \infty$, $f \in L_p(\mathbb{T})$, $\sigma = 1/p - 1/q$, $l, k \in \mathbb{N}$, and let*

$$\Omega_l(f;p;\sigma;q) \equiv \Big(\sum_{\nu=1}^{\infty} \nu^{q\sigma-1} \omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q} < \infty. \tag{1}$$

Then $f \in L_q(\mathbb{T})$ and the following estimates hold:

$$\|f\|_q \le C_1(l,p,q)\big\{\|f\|_1 + \Omega_l(f;p;\sigma;q)\big\}; \tag{2}$$

$$\omega_k\Big(f;\frac{\pi}{n}\Big)_q \le C_2(k,l,p,q)\Big(\sum_{\nu=n+1}^{\infty} \nu^{q\sigma-1} \omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q}, \quad n \in \mathbb{N}, \quad l \le k\,; \tag{3}$$

$$\omega_k\Bigl(f;\frac{\pi}{n}\Bigr)_q \le C_3(k,l,p,q)\biggl\{\biggl(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Bigl(f;\frac{\pi}{\nu}\Bigr)_p\biggr)^{1/q}+ \\ + n^{-k}\biggl(\sum_{\nu=1}^{n}\nu^{q(k+\sigma)-1}\omega_l^q\Bigl(f;\frac{\pi}{\nu}\Bigr)_p\biggr)^{1/q}\biggr\},\quad n\in\mathbb{N},\quad l>k. \tag{4}$$

Hereinafter, $C_j(k,l,p,q,\ldots)$, where $j\in\mathbb{N}$, stand for positive values depending only on the parameters given in parentheses.

Remark 1. In addition to the background outlined in [3], certain facts the author has learned after the publication of [3] should be mentioned, which partly provide more detailed information about the situation in the matter under consideration.

1) The first part of Theorem A: $\Omega_l(f;p;\sigma;q)<\infty \Rightarrow f\in L_q(\mathbb{T})$ and estimate (2) for $l=1$ were established by Ul'yanov [11, §3, Theorem 1, statement c, inequalities (3.6)].

2) The problem of establishing estimates of type (2) by methods different from those applied in [11] was also considered by Timan [9; 10, Theorem A]. In [10, first indention after the statement of Theorem A], it was noted that the first part of Theorem A above with estimate (2) (under the assumption that $\int_{\mathbb{T}} f(x)\,dx = 0$, which ensures the absence of the term $\|f\|_1$ on the right-hand side of (2); see [10, inequality (1.12)]) for $1<p<q\le 2$ was obtained in [8]. Actually, [8] (see [8, Theorem 8]) does not contain estimate (2); instead, there were announced an assertion that leads to the implication $\Omega_l(f;p;\sigma;\theta)<\infty \Rightarrow f\in L_q(\mathbb{T})$, where $\theta=\min\{2,p\}=p<q$.

3) Estimate (2) in various forms has also been obtained earlier by other authors (see, for example, [7, Introduction] and the references therein).

4) Estimate (3) for $l=k=1$ was proved by Ul'yanov [12, §4, Theorem 4, inequality (4.4)] (its formulation was given earlier in [11, §3, second inequality in (3.6′)], and the validity of this estimate for $l=k>1$ was also mentioned there).

5) Estimate (3) follows immediately from inequality (2) (see, for example, [7, Sect. 4], where this fact was noted for the case $l=k=1$). In the general case $l\le k$, it is sufficient to apply (2) to the function $\Delta_h^k f(x)$, where $h\in\mathbb{R}$, $|h|\le\pi/n$, and take into account the estimates $\omega_l(\Delta_h^k f;\pi/\nu)_p\le 2^k\omega_l(f;\pi/n)_p$ for $\nu\le n$ and $\omega_l(\Delta_h^k f;\pi/\nu)_p\le 2^k\omega_l(f;\pi/\nu)_p$ for $\nu\ge n+1$.

6) In [8, Theorem 8, inequality (40)], it was announced an inequality from which estimate (4) (with an additional term of order $O(n^{-k})$ on the right-hand side) can be obtained with $\theta=\min\{2,p\}<q$ instead of q on the right-hand side of this estimate.

Estimate (3) can be strengthened in the case $p>1$; more exactly, the following theorem holds.

Theorem B. *Suppose that* $1<p<q<\infty$, $f\in L_p(\mathbb{T})$, $\sigma=1/p-1/q$, $l,k\in\mathbb{N}$, $l\le k$, *and condition* (1) *holds. Then, the following estimate is valid*:

$$n^{\sigma-l}\biggl(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}\omega_k^p\Bigl(f;\frac{\pi}{\nu}\Bigr)_q\biggr)^{1/p}\le C_4(k,l,p,q)\biggl(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Bigl(f;\frac{\pi}{\nu}\Bigr)_p\biggr)^{1/q}. \tag{5}$$

Estimate (5) was first obtained by Kolyada [7, Sect. 3, Theorem 2, inequality (3.8)] for the case $l=k=1$; its validity for $l=k>1$ was noted by Goldman [2, Sect. 4, proof of Lemma 6, inequality (11)]. Estimate (5) for $k>l$ follows from the well-known order equality

$$\sum_{\nu=1}^{n}\nu^{\alpha\beta-1}\omega_k^\alpha(f;\pi/\nu)_q \asymp \sum_{\nu=1}^{n}\nu^{\alpha\beta-1}\omega_l^\alpha\Bigl(f;\frac{\pi}{\nu}\Bigr)_q,\quad n\in\mathbb{N}\cup\{+\infty\},$$

where $1\le\alpha<\infty$ and $0<\beta<\min\{k,l\}$.

Recall that an order equality $\varphi_n \asymp \psi_n$ means that there exist numbers $0 < C_5 \le C_6$ depending only on the parameters given (in this case, on k, l, β, and α) such that $C_5\psi_n \le \varphi_n \le C_6\psi_n$.

For given $p \in [1, \infty)$, denote by $M_p(\mathbb{T})$ the class of all functions $f \in L_p(\mathbb{T})$ whose Fourier coefficients satisfy the conditions $a_0(f) = 0$ and $a_n(f) \downarrow 0$ and $b_n(f) \downarrow 0$ as $n \uparrow \infty$. It is known (see, for example, [1, Ch. 1, Sect. 30]) that Fourier series of such functions converge everywhere expect maybe a countable set of points $x \equiv 0 \pmod{2\pi}$; i. e., we have

$$f(x) = \sum_{n=1}^{\infty}(a_n(f)\cos nx + b_n(f)\sin nx)$$

almost everywhere in $\mathbb{R}$.

In the present paper, which is a continuation of the author's research [5, 6], we consider the problem of optimality of inequalities (3), (4), and (5) in terms of order equalities on the whole class $M_p(T)$ for $1 < p < q < \infty$.

Theorem 1. *Let $1 < p < q < \infty$, $\sigma = 1/p - 1/q$, and $l, k \in \mathbb{N}$. A function $f \in M_p(\mathbb{T})$ belongs to $L_q(\mathbb{T})$ if and only if the condition $\Omega_l(f; p; \sigma; q) < \infty$ holds. Moreover, the following order equalities hold*:

$$\|f\|_q \asymp \Omega_l(f; p; \sigma; q) \equiv \left(\sum_{\nu=1}^{\infty} \nu^{q\sigma-1}\omega_l^q\Big(f; \frac{\pi}{\nu}\Big)_p\right)^{1/q}; \tag{6}$$

$$\omega_k\Big(f; \frac{\pi}{n}\Big)_q + n^\sigma\omega_l\Big(f; \frac{\pi}{n}\Big)_p \asymp \left(\sum_{\nu=n+1}^{\infty} \nu^{q\sigma-1}\omega_l^q\Big(f; \frac{\pi}{\nu}\Big)_p\right)^{1/q}, \quad n \in \mathbb{N}, \quad l \le k; \tag{7}$$

$$\begin{gathered}\omega_k\Big(f; \frac{\pi}{n}\Big)_q \asymp \left\{\left(\sum_{\nu=n+1}^{\infty} \nu^{q\sigma-1}\omega_l^q\Big(f; \frac{\pi}{\nu}\Big)_p\right)^{1/q} + \right.\\ \left. + n^{-k}\left(\sum_{\nu=1}^{n} \nu^{q(k+\sigma)-1}\omega_l^q\Big(f; \frac{\pi}{\nu}\Big)_p\right)^{1/q}\right\}, \quad n \in \mathbb{N}, \quad l > k;\end{gathered} \tag{8}$$

$$n^{-(l-\sigma)}\left(\sum_{\nu=1}^{n} \nu^{p(l-\sigma)-1}\omega_k^p\Big(f; \frac{\pi}{\nu}\Big)_q\right)^{1/p} \asymp \left(\sum_{\nu=n+1}^{\infty} \nu^{q\sigma-1}\omega_l^q\Big(f; \frac{\pi}{\nu}\Big)_p\right)^{1/q}, \quad n \in \mathbb{N}, \quad l \le k; \tag{9}$$

$$\omega_k\Big(f; \frac{\pi}{n}\Big)_q + n^\sigma\omega_l\Big(f; \frac{\pi}{n}\Big)_p \asymp n^{-(l-\sigma)}\left(\sum_{\nu=1}^{n} \nu^{p(l-\sigma)-1}\omega_k^p\Big(f; \frac{\pi}{\nu}\Big)_q\right)^{1/p}, \quad n \in \mathbb{N}, \quad l \le k. \tag{10}$$

Remark 2. When evaluating from below in the order equality (7) the second term $n^\sigma\omega_l(f; \pi/n)_p$ cannot be omitted in the general case, because there exists a function $g \in M_p(\mathbb{T})$ such that $n^\sigma\omega_l(g; \pi/n)_p \ne O(\omega_k(g; \pi/n)_q)$. The function $g \in M_p(\mathbb{T})$ is defined as follows (see [6, Sect. 3.1]): $g(x) = \sum\limits_{n=1}^{\infty} a_n \cos nx$, where $a_n = a_n(p; l) = n^{-(l+1-1/p)}$, $n \in \mathbb{N}$, and the following order equalities hold: $E_{n-1}(g)_p \asymp n^{-l}$, $n \in \mathbb{N}$, and $\omega_l(g; \pi/n)_p \asymp n^{-l}(\ln(en))^{1/p}$, $n \in \mathbb{N}$. Since $l > \sigma$, we have $g \in M_q(\mathbb{T})$; moreover,

$$E_{n-1}(g)_q \asymp n^{-(l-\sigma)}, \quad n \in \mathbb{N} \Leftrightarrow \omega_l(g; \pi/n)_q \asymp n^{-(l-\sigma)}, \quad n \in \mathbb{N}.$$

Thus, in view of these order equalities, we have

$$n^\sigma\omega_l(g; \pi/n)_p \asymp n^{-(l-\sigma)}(\ln(en))^{1/p} \asymp \omega_l(g; \pi/n)_q(\ln(en))^{1/p}, \quad n \in \mathbb{N};$$

whence $n^\sigma\omega_l(g; \pi/n)_p \ne O(\omega_l(g; \pi/n)_q)$, $n \in \mathbb{N}$, and *a fortiori* $n^\sigma\omega_l(g; \pi/n)_p \ne O(\omega_k(g; \pi/n)_q)$, $n \in \mathbb{N}$, in the case $k > l$.

However, it can be omitted if the sequence $\{\omega_l(f;\pi/n)_p\}_{n=1}^{\infty}$ satisfies Stechkin's (S_l)-condition $\big(\{\omega_l(f;\pi/n)_p\}_{n=1}^{\infty} \in S_l\big)$: there exists $\varepsilon \in (0,l)$ such that the sequence $\{n^{l-\varepsilon}\omega_l(f;\pi/n)_p\}_{n=1}^{\infty}$ almost increases. This condition is equivalent to Bari's $(B_l^{(\alpha)})$-condition for every fixed $\alpha \in [1,\infty)$ $\big(\big\{\omega_l\big(f;\pi/n\big)_p\big\}_{n=1}^{\infty} \in B_l^{(\alpha)}\big)$:

$$n^{-l}\bigg(\sum_{\nu=1}^{n}\nu^{\alpha l-1}\omega_l^{\alpha}\Big(f;\frac{\pi}{\nu}\Big)_p\bigg)^{1/\alpha} = O\Big(\omega_l\Big(f;\frac{\pi}{n}\Big)_p\Big), \quad n \in \mathbb{N},$$

(see [6, Sect. 3.2)].

Theorem 2. *Suppose that* $1 < p < q < \infty$, $f \in M_p(\mathbb{T})$, $\sigma = 1/p - 1/q$, $l,k \in \mathbb{N}$, $l \le k$, *and condition* (1) *holds. If* $\{\omega_l(f;\pi/n)_p\}_{n=1}^{\infty} \in S_l$, *then the following order equality holds*:

$$\omega_k\Big(f;\frac{\pi}{n}\Big)_q \asymp \bigg(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\bigg)^{1/q}, \quad n \in \mathbb{N}. \tag{11}$$

Remark 3. The condition $\{\omega_l(f;\pi/n)_p\}_{n=1}^{\infty} \in S_l$ guarantees the validity of the estimate $n^{\sigma}\omega_l\big(f;\pi/n\big)_p \le C_7(k,l,p,q)\omega_k\big(f;\pi/n\big)_q$, $n \in \mathbb{N}$, for all functions $f \in M_q(\mathbb{T})$, where $1 < p < q < \infty$, $l,k \in \mathbb{N}$ (see the proof of Theorem 2, inequality (22)). In the case $l > k$, this estimate holds for $f \in M_q(\mathbb{T})$ without any conditions on the sequence $\{\omega_l(f;\pi/n)_p\}_{n=1}^{\infty}$ (see the proof of Theorem 1, inequality (18)).

Remark 4. In connection with Theorem 2, note also the following fact, which is an obvious corollary of the order equality (9) in Theorem 1: *the order equality* (11) *is valid if and only if* $\big\{\omega_k(f;\pi/n)_q\big\}_{n=1}^{\infty} \in B_{l-\sigma}^{(p)}$. In addition, if $\big\{\omega_l(f;\pi/n)_p\big\}_{n=1}^{\infty} \in B_l^{(p)} \Leftrightarrow \big\{\omega_l(f;\pi/n)_p\big\}_{n=1}^{\infty} \in S_l$, then, in view of (11) and (9), we have $\{\omega_k(f;\pi/n)_q\}_{n=1}^{\infty} \in B_{l-\sigma}^{(p)}$. On the other hand, in view of (11), the latter condition guarantees only that $\big\{\big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q(f;\pi/\nu)_p\big)^{1/q}\big\}_{n=1}^{\infty} \in B_{l-\sigma}^{(p)}$ but does not that $\big\{\omega_l(f;\pi/n)_p\big\}_{n=1}^{\infty} \in B_l^{(p)}$.

P r o o f of Theorem 1. In the proof of Theorem 1, we will use neither estimates (2)–(5) from Theorems A and B nor the direct (see [6, inequality (0.3)]) and inverse (see [5, inequality (2)]) theorems of approximation theory for periodic functions in different metrics. We will use only certain known results that are characteristic of functions from the class $M_p(\mathbb{T})$ (the corresponding statements are gathered in [6, Sect. 1]). Since the auxiliary inequalities needed for the proofs of Theorems 1 and 2 were established by the author in [5, 6], we will mostly refer to these papers instead of giving original references, which can be found in [5] and [6].

1) The first part of Theorem 1 and the order equality (6) were proved in [6, Sect. 2, the proof of statement (1) of Theorem 1]; more precisely, if $f \in M_p(\mathbb{T})$ and $\Omega_l(f;p;\sigma;q) < \infty$, then $f \in L_q(\mathbb{T})$ and $\|f\|_q \le C_8(l,p,q)\Omega_l(f;p;\sigma;q)$ (the sufficiency); if $f \in M_p(\mathbb{T})$ belongs to $L_q(\mathbb{T})$, then $f \in M_q(\mathbb{T})$ and $\Omega_l(f;p;\sigma;q) \le C_9(l,p,q)\|f\|_q$ (the necessity). Moreover, in the proof of statement (1) of [6, Theorem 1], we actually obtained the following estimates for a function $f \in M_p(\mathbb{T})$ under the condition $\Omega_l(f;p;\sigma;q) < \infty$ (see also [5, Sect. 1]):

$$C_{10}(p,q)E(f;p;\sigma;q) \le \|f\|_q \le C_{11}(p,q)E(f;p;\sigma;q), \tag{12}$$

where

$$E(f;p;\sigma;q) \equiv \bigg(\sum_{\nu=1}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^q(f)_p\bigg)^{1/q} \asymp \bigg(\sum_{\nu=1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\bigg)^{1/q} \equiv \Omega_l(f;p;\sigma;q).$$

2) *The upper estimate in* (7): taking into account the inequality (see [6, Sect. 2, inequality (2.1)])

$$n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p \le C_{12}(l,p,q)\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q}, \quad n\in\mathbb{N}, \quad l\in\mathbb{N}, \tag{13}$$

in the estimation of $E_{n-1}(f)_q$ from above (see [6, Sect. 2, step *The upper estimate* in the proof of statement (2) of Theorem 1]), we obtain

$$E_{n-1}(f)_q \le C_{13}(l,p,q)\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q}, \quad n\in\mathbb{N}, \quad l\in\mathbb{N}. \tag{14}$$

Further, using estimate (14) in the inequality (see [6, Sect. 1, Lemma 1, inequality (1.8)])

$$\omega_k\Big(f;\frac{\pi}{n}\Big)_q \le C_{14}(k,p,q)\Big\{E_n(f)_q + n^{\sigma}\omega_k\Big(f;\frac{\pi}{n}\Big)_p\Big\}, \quad n\in\mathbb{N}, \quad k\in\mathbb{N},$$

and taking into account (13), we obtain for $l\le k$

$$\begin{aligned}\omega_k\Big(f;\frac{\pi}{n}\Big)_q &\le C_{14}(k,p,q)\Big\{C_{13}(l,p,q)\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q} + 2^{k-l}n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p\Big\} \le \\ &\le C_{15}(k,l,p,q)\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q}, \quad n\in\mathbb{N},\end{aligned} \tag{15}$$

where $C_{15}(k,l,p,q) = C_{14}(k,p,q)\{C_{13}(l,p,q) + 2^{k-l}C_{12}(l,p,q)\}$; whence,

$$\omega_k\Big(f;\frac{\pi}{n}\Big)_q + n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p \le (C_{15}(k,l,p,q)+C_{12}(l,p,q))\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q}, \quad n\in\mathbb{N}.$$

The lower estimate in (7): applying to the inequality (see [6, Sect. 2, step *The lower estimate* in the proof of statement (2) of Theorem 1])

$$\begin{aligned}&\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q} \le \\ &\le C_{16}(l,p,q)\Big\{n^{\sigma-l}\Big(\sum_{\nu=1}^{n}\nu^{pl-1}E_{\nu-1}^p(f)_p\Big)^{1/p} + \Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^q(f)_p\Big)^{1/q}\Big\}\end{aligned}$$

the lower estimate in the order equality (1.7) from [6, Sect. 1, Proposition 5]:

$$\omega_l\Big(f;\frac{\pi}{n}\Big)_p \asymp n^{-l}\Big(\sum_{\nu=1}^{n}\nu^{pl-1}E_{\nu-1}^p(f)_p\Big)^{1/p}, \quad n\in\mathbb{N}, \quad l\in\mathbb{N}, \tag{16}$$

and inequality (10) from [5, Sect. 1]:

$$\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^q(f)_p\Big)^{1/q} \le C_{17}(k,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q, \quad n\in\mathbb{N}, \quad k\in\mathbb{N},$$

we obtain for arbitrary $l,k\in\mathbb{N}$

$$\begin{aligned}\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q} &\le C_{16}(l,p,q)\Big\{C_{18}(l,p)n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p + \\ + C_{17}(k,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q\Big\} &\le C_{19}(k,l,p,q)\Big\{\omega_k\Big(f;\frac{\pi}{n}\Big)_q + n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p\Big\}, \quad n\in\mathbb{N}.\end{aligned} \tag{17}$$

3) *The upper estimate in* (8): if $f \in M_p(\mathbb{T})$ and $\Omega_l(f;p;\sigma;q) < \infty$, then $f \in L_q(\mathbb{T})$ and, consequently, $f \in M_q(\mathbb{T})$. By the upper estimate in (16) and inequality (14), we have

$$\omega_k\Big(f;\frac{\pi}{n}\Big)_q \le C_{20}(k,q)n^{-k}\bigg(\sum_{\nu=1}^{n}\nu^{qk-1}E^q_{\nu-1}(f)_q\bigg)^{1/q} \le$$

$$\le C_{20}(k,q)C_{13}(l,p,q)n^{-k}\bigg(\sum_{\nu=1}^{n}\nu^{qk-1}\sum_{\mu=\nu+1}^{n}\mu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p +$$

$$+\sum_{\nu=1}^{n}\nu^{qk-1}\sum_{\mu=n+1}^{\infty}\mu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p\bigg)^{1/q} \le$$

$$\le C_{21}(k,l,p,q)n^{-k}\bigg(\sum_{\mu=1}^{n}\mu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p\sum_{\nu=1}^{\mu}\nu^{qk-1}+\sum_{\nu=1}^{n}\nu^{qk-1}\sum_{\mu=n+1}^{\infty}\mu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p\bigg)^{1/q} \le$$

$$\le C_{21}(k,l,p,q)n^{-k}\bigg(\sum_{\mu=1}^{n}\mu^{q(k+\sigma)-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p + n^{qk}\sum_{\mu=n+1}^{\infty}\mu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p\bigg)^{1/q} \le$$

$$\le C_{21}(k,l,p,q)\bigg\{n^{-k}\bigg(\sum_{\mu=1}^{n}\mu^{q(k+\sigma)-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p\bigg)^{1/q} + \bigg(\sum_{\mu=n+1}^{\infty}\mu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\mu}\Big)_p\bigg)^{1/q}\bigg\}, \quad n\in\mathbb{N}.$$

The lower estimate in (8): let us first prove that the following estimate holds for $l > k$:

$$n^\sigma\omega_l\Big(f;\frac{\pi}{n}\Big)_p \le C_{22}(k,l,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q, \quad n\in\mathbb{N}. \tag{18}$$

To this end, we will need the inequality (see [6, Sect. 2, inequality (2.6)]

$$n^\sigma E_{n-1}(f)_p \le C_{23}(k,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q, \quad n\in\mathbb{N}, \quad k\in\mathbb{N}. \tag{19}$$

Applying (19) in the upper estimate in (16) and taking into account the following known property of the modulus of smoothness: $\delta_2^{-k}\omega_k(f;\delta_2)_q \le 2^k\delta_1^{-k}\omega_k(f;\delta_1)_q$ for $0<\delta_1\le\delta_2 \Leftrightarrow \nu^k\omega_k(f;\pi/\nu)_q \le 2^kn^k\omega_k(f;\pi/n)_q$ for $1\le\nu\le n$, we obtain

$$n^\sigma\omega_l\Big(f;\frac{\pi}{n}\Big)_p \le C_{20}(l,p)n^{\sigma-l}\bigg(\sum_{\nu=1}^{n}\nu^{pl-1}E^p_{\nu-1}(f)_p\bigg)^{1/p} \le$$

$$\le C_{20}(l,p)C_{23}(k,p,q)n^{\sigma-l}\bigg(\sum_{\nu=1}^{n}\nu^{pl-1-p\sigma}\omega_k^p\Big(f;\frac{\pi}{\nu}\Big)_q\bigg)^{1/p} \le$$

$$\le C_{24}(k,l,p,q)n^{\sigma-l}2^kn^k\omega_k\Big(f;\frac{\pi}{n}\Big)_q\bigg(\sum_{\nu=1}^{n}\nu^{p(l-k-\sigma)-1}\bigg)^{1/p} \le C_{25}(k,l,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q,$$

which implies the required estimate in (18).

Applying estimate (18) in inequality (17), we obtain the following upper estimate for the first term on the right-hand side of (8):

$$\bigg(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\bigg)^{1/q} \le C_{19}(k,l,p,q)(1+C_{22}(k,l,p,q))\omega_k\Big(f;\frac{\pi}{n}\Big)_q, \quad n\in\mathbb{N}.$$

The following estimate was obtained in [5, Sect. 1, the proof of the lower estimate for the second term on the right-hand side of the order equality (7)]:

$$n^{-k}\Big(\sum_{\nu=1}^{n}\nu^{q(k+\sigma)-1}E_{\nu-1}^{q}(f)_p\Big)^{1/q}\le C_{26}(k,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q,\quad n\in\mathbb{N}.$$

Hence, in view of the known order equality (see, for example, [4, Sect. 2, Remark 7, order equality (15)])

$$\sum_{\nu=1}^{n}\nu^{\alpha\beta-1}E_{\nu-1}^{\alpha}(f)_p\asymp\sum_{\nu=1}^{n}\nu^{\alpha\beta-1}\omega_l^{\alpha}\Big(f;\frac{\pi}{\nu}\Big)_p, \tag{20}$$

where $n\in\mathbb{N}\cup\{+\infty\}$, $1\le\alpha<\infty$, and $0<\beta<l$, we obtain ($l>k\Rightarrow l>k+\sigma$, $\sigma\in(0,1)$)

$$n^{-k}\Big(\sum_{\nu=1}^{n}\nu^{q(k+\sigma)-1}\omega_l^{q}\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q}\asymp n^{-k}\Big(\sum_{\nu=1}^{n}\nu^{q(k+\sigma)-1}E_{\nu-1}^{q}(f)_p\Big)^{1/q}\le$$

$$\le C_{26}(k,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q,\quad n\in\mathbb{N}.$$

The latter inequality implies the upper estimate for the second term on the right-hand side of (8).

4) *The upper estimate in* (9): the upper estimate in (12) implies the inequality

$$E_{n-1}(f)_q\le C_{27}(p,q)\Big\{n^{\sigma}E_{n-1}(f)_p+\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^{q}(f)_p\Big)^{1/q}\Big\},\quad n\in\mathbb{N}. \tag{21}$$

Indeed, applying (12) to the function $g_n(f;x)=f(x)-S_n(f;x)$, where $S_n(f;x)$ is the partial sum of order $n\in\mathbb{N}$ of the Fourier series of the function $f\in M_p(T)$, and taking into account the estimate $E_{\nu-1}(g_n)_p\le\|g_n\|_p\le(1+C_{28}(p))E_n(f)_p$, $\nu\in\mathbb{N}$, where $C_{28}(p)$ is the constant in the well-known M. Riesz inequality $\|S_n(f;\cdot)\|_p\le C_{28}(p)\|f\|_p$ ($1<p<\infty$, $f\in L_p(\mathbb{T})$), and the equality $E_{\nu-1}(g_n)_p=E_{\nu-1}(f)_p$, $\nu\ge n+1$, $n\in\mathbb{Z}_+$, we obtain for $n\in\mathbb{N}$

$$E_n(f)_q\le\|g_n(f;\cdot)\|_q\le C_{11}(p,q)E(g_n;p;\sigma;q)\le$$

$$\le C_{11}(p,q)\Big\{\Big(\sum_{\nu=1}^{n+1}\nu^{q\sigma-1}E_{\nu-1}^{q}(g_n)_p\Big)^{1/q}+\Big(\sum_{\nu=n+2}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^{q}(g_n)_p\Big)^{1/q}\Big\}\le$$

$$\le C_{11}(p,q)\Big\{(1+C_{28}(p))E_n(f)_p\Big(\sum_{\nu=1}^{n+1}\nu^{q\sigma-1}\Big)^{1/q}+\Big(\sum_{\nu=n+2}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^{q}(f)_p\Big)^{1/q}\Big\}\le$$

$$\le C_{11}(p,q)\Big\{(1+C_{28}(p))C_{29}(q,\sigma)(n+1)^{\sigma}E_n(f)_p+\Big(\sum_{\nu=n+2}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^{q}(f)_p\Big)^{1/q}\Big\},$$

and for $n=0$

$$E_0(f)_q\le\|f\|_q\le C_{11}(p,q)E(f;p;\sigma;q)\le C_{11}(p,q)\Big\{E_0(f)_p+\Big(\sum_{\nu=2}^{\infty}\nu^{q\sigma-1}E_{\nu-1}^{q}(f)_p\Big)^{1/q}\Big\}.$$

Inequality (21) was used in [6, Sect. 2, step *The upper estimate* in the proof of statement (3) of Theorem 1] for obtaining the estimate

$$n^{\sigma-l}\Big(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}E_{\nu-1}^{p}(f)_q\Big)^{1/p}\le C_{30}(l,p,q)\Big(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^{q}\Big(f;\frac{\pi}{\nu}\Big)_p\Big)^{1/q},\quad n\in\mathbb{N};$$

whence, by the order equality (20), we obtain the following upper estimate in (9):

$$n^{-(l-\sigma)}\left(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}\omega_k^p\Big(f;\frac{\pi}{\nu}\Big)_q\right)^{1/p}\le C_{31}(k,l,p,q)\left(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\right)^{1/q},\quad n\in\mathbb{N},\quad l\le k;$$

The lower estimate in (9): by inequality (17), we have for all $l,k\in\mathbb{N}$

$$\left(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\right)^{1/q}\le C_{19}(k,l,p,q)\Big\{\omega_k\Big(f;\frac{\pi}{n}\Big)_q+n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p\Big\},\quad n\in\mathbb{N}.$$

The upper estimate for the first term $\omega_k(f;\pi/n)_q$ on the right-hand side is obvious, because, in view of the fact that $\omega_k(f;\pi/n)_q\downarrow(n\uparrow)$, we have

$$n^{-(l-\sigma)}\left(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}\omega_k^p\Big(f;\frac{\pi}{\nu}\Big)_q\right)^{1/p}\ge$$

$$\ge n^{-(l-\sigma)}\omega_k\Big(f;\frac{\pi}{n}\Big)_q\left(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}\right)^{1/p}\ge C_{32}(l,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q.$$

The following upper estimate for the second term $n^{\sigma}\omega_l(f;\pi/n)_p$ was established above (see the proof of the lower estimate in (8) at step 3):

$$n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p\le C_{20}(l,p)C_{23}(k,p,q)n^{\sigma-l}\left(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}\omega_k^p\Big(f;\frac{\pi}{\nu}\Big)_q\right)^{1/p}.$$

Combining the obtained inequalities, we come to the required lower estimate in the order equality (9):

$$\left(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\right)^{1/q}\le C_{33}(k,l,p,q)n^{-(l-\sigma)}\left(\sum_{\nu=1}^{n}\nu^{p(l-\sigma)-1}\omega_k^p\Big(f;\frac{\pi}{\nu}\Big)_q\right)^{1/p},\quad n\in\mathbb{N}.$$

5) The order equality (10) follows from (7) and (9). The proof of Theorem 1 is complete. □

P r o o f of Theorem 2. The upper estimate in (11) was obtained at step 2 of the proof of Theorem 1 (see inequality (15)):

$$\omega_k\Big(f;\frac{\pi}{n}\Big)_q\le C_{15}(k,l,p,q)\left(\sum_{\nu=n+1}^{\infty}\nu^{q\sigma-1}\omega_l^q\Big(f;\frac{\pi}{\nu}\Big)_p\right)^{1/q},\quad n\in\mathbb{N}.$$

To obtain the lower estimate in (11), we preliminarily prove that, if $\{\omega_l(f;\pi/n)_p\}_{n=1}^{\infty}\in S_l$, then the following estimate holds for all $l,k\in\mathbb{N}$:

$$n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p\le C_{34}(k,l,p,q)\omega_k\Big(f;\frac{\pi}{n}\Big)_q,\quad n\in\mathbb{N}. \tag{22}$$

Indeed, since

$$\Big\{\omega_l\Big(f;\frac{\pi}{n}\Big)_p\Big\}_{n=1}^{\infty}\in S_l\Leftrightarrow\Big\{\omega_l\Big(f;\frac{\pi}{n}\Big)_p\Big\}_{n=1}^{\infty}\in B_l^{(p)}\Leftrightarrow\{E_{n-1}(f)_p\}_{n=1}^{\infty}\in B_l^{(p)}$$

(see [6, Sect. 3.2)]), in view of [6, Sect. 2, inequality (2.8)] and [6, Introduction, inequality (0.4)], we obtain

$$n^{\sigma}\omega_l\Big(f;\frac{\pi}{n}\Big)_p\le C_{35}(l,p,q)E_n(f)_q\le C_{35}(l,p,q)C_{36}(k)\omega_k\Big(f;\frac{\pi}{n}\Big)_q,\quad n\in\mathbb{N}.$$

The required lower estimate in (11) follows from (17) and (22):

$$\left(\sum_{\nu=n+1}^{\infty} \nu^{q\sigma-1}\omega_l^q\left(f;\frac{\pi}{\nu}\right)_p\right)^{1/q} \le C_{19}(k,l,p,q)\big(1+C_{34}(k,l,p,q)\big)\omega_k\left(f;\frac{\pi}{n}\right)_q, \quad n\in\mathbb{N}.$$

The proof of Theorem 2 is complete. □

Remark 5. By inequality (19), the upper estimate in the order equality (16) implies the inequality

$$\omega_l\left(f;\frac{\pi}{n}\right)_p \le C_{37}(k,l,p,q)n^{-l}\left(\sum_{\nu=1}^{n} \nu^{p(l-\sigma)-1}\omega_k^p\left(f;\frac{\pi}{\nu}\right)_q\right)^{1/p}, \quad l,k,n\in\mathbb{N}. \tag{23}$$

Inequality (23) (the case $l \le k$) and inequality (18) (the case $l > k$) for functions $f \in M_q(\mathbb{T}) \subset M_p(\mathbb{T})$ are inverse (in the sense of the upper estimate for $\omega_l(f;\delta)_p$ in terms of $\omega_k(f;\delta)_q$) to inequalities (3) and (4), respectively, which hold for all functions $f \in L_q(\mathbb{T})$ under the condition of convergence of series in (1). From inequalities (23) and (18), we can conclude that, in the passage from the class $M_q(\mathbb{T})$ to the class $M_p(\mathbb{T})$, where $p < q$, the smoothness of a function $f \in M_q(\mathbb{T})$ increases by a value not larger than σ in the case $l \le k$ (see [6, Sect. 3.3)], where the author considered the case $k = l$ and $\omega_l(f;\delta)_q \asymp \delta^\alpha$, $0 < \alpha \le l$, $\delta \in (0,\pi]$) and increases by a value not smaller than σ in the case $l > k$.

REFERENCES

1. Bary N. K. *A Treatise on Trigonometric Series. Vols.* I, II. Oxford, New York: Pergamon Press, 1964, Vol. I, 533 p; Vol. II, 508 p. Original Russian text published in *Trigonometricheskie ryady*, Moscow: Fiz.-Mat. Giz. Publ., 1961, 936 p.
2. Gol'dman M. L. An imbedding criterion for different metrics for isotropic Besov spaces with arbitrary moduli of continuity. *Proc. Steklov Inst. Math.*, 1994. No. 2. P. 155–181.
3. Il'yasov N. A. On the inequality between modulus of smoothness of various orders in different metrics. *Math. Notes*, 1991. Vol. 50, No. 2. P. 877–879. DOI: 10.1007/BF01157580
4. Il'yasov N. A. On the direct theorem of approximation theory of periodic functions in different metrics. *Proc. Steklov Inst. Math.*, 1997. Vol. 219. P. 215–230.
5. Il'yasov N. A. The inverse theorem in various metrics of approximation theory for periodic functions with monotone Fourier coefficients. *Trudy Inst. Mat. i Mekh. UrO RAN* [Proc. of Krasovskii Institute of Mathematics and Mechanics of the UB RAS], 2016. Vol. 22, No. 4. P. 153–162. (in Russian) DOI: 10.21538/0134-4889-2016-22-4-153-162
6. Il'yasov N. A. The direct theorem of the theory of approximation of periodic functions with monotone Fourier coefficients in different metrics. *Proc. Steklov Inst. Math.*, 2018. Vol. 303, Suppl. 1. P. S92–S106. DOI: 10.1134/S0081543818090109
7. Kolyada V. I. On relations between moduli of continuity in different metrics. *Proc. Steklov Inst. Math.*, 1989. Vol. 181. P. 127–148.
8. Timan M. F. Best approximation and modulus of smoothness of functions defined on the entire real axis. *Izv. Vyssh. Ucheb. Zaved. Mat.*, 1961. No. 6. P. 108–120. (in Russian)
9. Timan M. F. Some embedding theorems for L_p-classes of functions. *Dokl. Akad. Nauk SSSR*,1970. Vol. 193, No. 6, P. 1251–1254. (in Russian)
10. Timan M. F. The imbedding of the $L_p^{(k)}$-classes of functions. *Izv. Vyssh. Ucheb. Zaved. Mat.*, 1974. No. 10(149). P. 61–74. (in Russian)
11. Ul'yanov P. L. The imbedding of certain function classes H_p^ω. *Math. USSR–Izv.*, 1968. Vol. 2, No. 3. P. 601–637.
12. Ul'yanov P. L. Imbedding theorems and relations between best approximations (moduli of continuity) in different metrics. *Math. USSR–Sb.*, 1970. Vol. 10, No. 1. P. 103–126.

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 33–42

DOI: 10.15826/umj.2018.2.005

SOME PROPERTIES OF OPERATOR EXPONENT

Lyudmila F. Korkina[†] **and Mark A. Rekant**[††]

Ural Federal University,
51 Lenin aven., Ekaterinburg, Russia, 620000

[†]l.f.korkina@urfu.ru, [††]m.a.rekant@urfu.ru

Abstract: We study operators given by series, in particular, operators of the form $e^B = \sum\limits_{n=0}^{\infty} B^n/n!$, where B is an operator acting in a Banach space X. A corresponding example is provided. In our future research, we will use these operators for introducing and studying functions of operators constructed (with the use of the Cauchy integral formula) on the basis of scalar functions and admitting a faster than power growth at infinity.

Keywords: Closed operator, Operator exponent, Multiplicative property.

The theory of functions of normal operators has been developed in Hilbert spaces [8, Ch. 12,13]. However, functions of an operator in Banach spaces are introduced under quite serious restrictions on the operator and the corresponding scalar functions (see e.g. [2, Ch. VII.3]). For a considerable class of operators, these scalar functions are assumed to be analytical with polynomial growth at infinity (see e.g. [1] and [6, Ch. 1, § 5]). The authors' papers [3–5] are in the same vein. In these papers, based on the Cauchy integral formula, functions of an operator were constructed in terms of natural powers of the operator. To introduce and study functions of an operator built constructed on the basis of scalar functions and admitting the growth at infinity faster than the power function but not faster than the exponential function have, we will need operators of the form

$$e^B = \sum_{n=0}^{\infty} \frac{B^n}{n!}, \tag{1}$$

where B is an operator on a Banach space X. In this paper, we study the properties of such operators.

We will use series of elements of a Banach space X and operator series. The principal notions of numerical series (double series and repeated series) are naturally extended to series of elements of the space X [7, Ch. 2, § 2]. In this paper, the convergence of partial sums of series from X is interpreted as the convergence in the norm of this space. For a series $\sum\limits_{n=0}^{\infty} A_n$ of operators A_n acting in X, its sum is the operator A with the domain $\mathcal{D}(A) = \Big\{x \in X\colon \sum\limits_{n=0}^{\infty} A_n x \text{ converges}\Big\}$ and such that $Ax = \sum\limits_{n=0}^{\infty} A_n x$ for $x \in \mathcal{D}(A)$. The expression $A \subset B$ $(B \supset A)$ for operators A and B means that B is the extension of A [7, Ch. 7, Sect. 6].

Let us proceed to the results.

In what follows, we will need the following auxiliary assertion.

Assertion 1. *The following statements hold*:

(i) (An analog of Abel's test for numerical series). Suppose that a series $\sum\limits_{n=0}^{\infty} a_n$ *converges in* X *and a sequence* $\{\alpha_n\}_{n=0}^{\infty} \subset \mathbb{R}$ *is monotonic and bounded. Then, the series* $\sum\limits_{n=0}^{\infty} \alpha_n a_n$ *converges.*

(ii) Suppose that $\{a_{m,n}\}_{m,n=0}^{\infty} \subset X$ *and the series* $\sum\limits_{m,n=0}^{\infty} a_{m,n}$ *converges absolutely. Then, every rearrangement of this series converges absolutely to the same sum.*

(iii) Suppose that the terms of a series $\sum\limits_{m,n=0}^{\infty} a_{m,n}$ $(a_{m,n} \in X)$ *are reindexed (with a single index) and the series* $\sum\limits_{k=0}^{\infty} b_k$ *is composed of them. If one of these two series or the repeated series* $\sum\limits_{m=0}^{\infty}\sum\limits_{n=0}^{\infty} a_{m,n}$ *converges absolutely, then the other two series converge absolutely to the same sum.*

(iv) If a series $\sum\limits_{m,n=0}^{\infty} a_{m,n}$ $(a_{m,n} \in X)$ *converges absolutely, then the series* $\sum\limits_{k=0}^{\infty} \sum\limits_{\substack{m+n=k \\ (m,n\geq 0)}}^{\infty} a_{m,n}$ *also converges absolutely to the same sum.*

The proof of statement (i) is almost the same as the proof of Abel's test for numerical series. The proofs of statements (ii)–(iv) reduce to the use of the corresponding statements for numerical series after the application a continuous linear functional to the series under consideration. Here, we take into account the fact that if values of all such functionals coincide at two elements from X, then these elements are equal [2, Ch. II.3.15].

Assertion 2. *Suppose that A is an operator acting in X, $x \in X$, $k \in \mathbb{N}$, a sequence $\{\alpha_n\}_{n=0}^{\infty} \subset \mathbb{R}$ is such that the sequence* $\left\{\dfrac{\alpha_{n+k}}{\alpha_n}\right\}$ *is monotonic and bounded, and the series* $\sum\limits_{n=0}^{\infty} \alpha_n A^{n+k}x$ *converges. Then, the series* $\sum\limits_{n=0}^{\infty} \alpha_n A^n x$ *converges. If the operator A^k is linear and closed, then the following equality holds:*

$$\sum_{n=0}^{\infty} \alpha_n A^{n+k}x = A^k \sum_{n=0}^{\infty} \alpha_n A^n x, \tag{2}$$

which is equivalent to the expression

$$\sum_{n=0}^{\infty} \alpha_n A^{n+k} \subset A^k \sum_{n=0}^{\infty} \alpha_n A^n. \tag{3}$$

P r o o f. The following relations are valid:

$$\sum_{n=0}^{\infty} \alpha_n A^n x = \sum_{n=0}^{k-1} \alpha_n A^n x + \sum_{n=0}^{\infty} \alpha_{n+k} A^{n+k} x = \sum_{n=0}^{k-1} \alpha_n A^n x + \sum_{n=0}^{\infty} \frac{\alpha_{n+k}}{\alpha_n} (\alpha_n A^{n+k} x).$$

The latter series converges by the analog of Abel's test. Moreover, under the assumption that the operator A^k is linear and closed, equality (2) holds. □

Remark 1. The operator A^k is linear and closed if the operator A is linear and its resolvent set $\rho(A) \neq \varnothing$ [2, VII.9.7].

Remark 2. The boundedness and monotonicity of the sequence $\left\{\dfrac{\alpha_{n+k}}{\alpha_n}\right\}$ starting from a certain index follow from the fact that the sequence $\left\{\dfrac{\alpha_{n+1}}{\alpha_n}\right\}$ is monotonic and bounded.

This remark follows from the relation

$$\frac{\alpha_{n+k}}{\alpha_n} = \frac{\alpha_{n+1}}{\alpha_n} \times \frac{\alpha_{n+2}}{\alpha_{n+1}} \times \cdots \times \frac{\alpha_{n+k}}{\alpha_{n+k-1}}$$

and the fact that terms of a monotonic sequence of reals are of the same sign starting from a certain index.

Remark 3. Equality (2) holds without the assumption that the sequence $\left\{\frac{\alpha_{n+k}}{\alpha_n}\right\}$ is monotonic and bounded if the operator A^k is linear and closed and both the series in (2) converge.

Corollary 1. *Suppose that $k \in \mathbb{N}$ and the operator A^k is linear and closed. Then,*

$$e^A A^k \subset A^k e^A.$$

To prove this fact, it is sufficient to take $\alpha_n = 1/n!$ in (3).

In Assertion 1, the sequence $\left\{\frac{\alpha_{n+k}}{\alpha_n}\right\}_{n=0}^{\infty}$ is required to be monotonic and bounded. Let us consider the conditions related to these properties.

Lemma 1. *If $a \in \mathbb{R}$ and a sequence $\{\alpha_n\} \subset (0, +\infty)$ is such that $\frac{\alpha_{n+1}}{\alpha_n} \le a$ for all n, then $\alpha_n \le Ca^n$ for all n, where $C = \alpha_1/a$. Conversely, if the sequence $\left\{\frac{\alpha_{n+1}}{\alpha_n}\right\}$ is monotonic and $\alpha_n \le Ca^n$ for some $C, a \in (0, +\infty)$ and all n, then the sequence $\left\{\frac{\alpha_{n+1}}{\alpha_n}\right\}$ is bounded.*

P r o o f. Suppose that $a \in \mathbb{R}$ is such that $\frac{\alpha_{n+1}}{\alpha_n} \le a$ for all n. Then,

$$\frac{\alpha_n}{\alpha_1} = \frac{\alpha_2}{\alpha_1} \times \frac{\alpha_3}{\alpha_2} \times \cdots \times \frac{\alpha_n}{\alpha_{n-1}} \le a^{n-1};$$

i. e., $\alpha_n \le Ca^n$ for $C = \alpha_1/a$.

Conversely, suppose that $\alpha_n \le Ca^n$ for some $C, a \in (0, +\infty)$, and all n and the sequence $\left\{\frac{\alpha_{n+1}}{\alpha_n}\right\}$ is monotonic. Denote by d the limit of this sequence, $d \in [0, +\infty]$. Assume that $d = +\infty$. Then $\lim\limits_{n\to\infty} \sqrt[n]{\alpha_n} = +\infty$. This contradicts the inequality $\sqrt[n]{\alpha_n} \le \sqrt[n]{C}a$. Therefore, $d \in \mathbb{R}$; i. e., the sequence $\left\{\frac{\alpha_{n+1}}{\alpha_n}\right\}$ is bounded. The lemma is proved. □

Note that the requirement of monotonicity in the second part of the lemma is essential.

Example 1. Suppose that a sequence $\{n_m\} \subset \mathbb{N}$ is such that $n_{m+1} > n_m + 1$, $2^{n_m} > m!$, and $\alpha_n = (m-1)!$ for $n_{m-1} < n \le n_m$ $(n_0 = 0)$ for all $m \in \mathbb{N}$. In this case, the sequence $\left\{\frac{\alpha_{n+1}}{\alpha_n}\right\}$ is unbounded, although $\alpha_n < 2^n$. (Indeed, if $n_{m-1} < n \le n_m$, then $\alpha_n = (m-1)! < 2^{n_{m-1}} < 2^n$).

Assertion 3. *Suppose that $k \in \mathbb{N}$ and $\{\alpha_n\}_{n=0}^{\infty} \subset (0, +\infty)$. If the sequence $\left\{\frac{\alpha_{n+k}}{\alpha_n}\right\}$ is bounded, then $\alpha_n \le Cb^n$ for some $b, C \in (0, +\infty)$ and all n. Conversely, if the sequence $\left\{\frac{\alpha_{n+k}}{\alpha_n}\right\}$ is monotonic, $C, b \in (0, +\infty)$, and $\alpha_n \le Cb^n$ for all n, then the sequence $\left\{\frac{\alpha_{n+k}}{\alpha_n}\right\}$ is bounded.*

P r o o f. Suppose that $a \in \mathbb{R}$ is such that $\frac{\alpha_{n+k}}{\alpha_n} \le a$ for all n. Let us consider the subsequences $\{\beta_m^{(r)}\}_{m=0}^{\infty}$ of $\{\alpha_n\}$ with $\beta_m^{(r)} = \alpha_{mk+r}$ $(r = 0, 1, \ldots, k-1)$. For all m, we have

$$\frac{\beta_{m+1}^{(r)}}{\beta_m^{(r)}} = \frac{\alpha_{(m+1)k+r}}{\alpha_{mk+r}} = \frac{\alpha_{(mk+r)+k}}{\alpha_{mk+r}} \le a.$$

Then, according to Lemma 1, there is a number $C_r \in (0, +\infty)$ such that

$$\beta_m^{(r)} \le C_r a^m = \frac{C_r}{a^{r/k}} (\sqrt[k]{a})^{mk+r}$$

for all m. Setting $C = \max\limits_{0 \le r \le k-1} \dfrac{C_r}{a^{r/k}}$ and $b = \sqrt[k]{a}$, we obtain $\alpha_n \le C b^n$ for all n.

Conversely, suppose that the sequence $\left\{\dfrac{\alpha_{n+k}}{\alpha_n}\right\}$ is monotonic, $C, b \in (0, +\infty)$, and $\alpha_n \le C b^n$ for all n. If $\lim\limits_{n\to\infty} \dfrac{\alpha_{n+k}}{\alpha_n}$ is finite, then there is nothing to prove. Assume that $\lim\limits_{n\to\infty} \dfrac{\alpha_{n+k}}{\alpha_n} = +\infty$. Again, introducing $\beta_m^{(r)} = \alpha_{mk+r}$ $(r = 0, 1, \ldots, k-1)$, we conclude that the sequence $\{\beta_m^{(r)}\}_{m=0}^{\infty}$ is monotonic because

$$\frac{\beta_{m+1}^{(r)}}{\beta_m^{(r)}} = \frac{\alpha_{(mk+r)+k}}{\alpha_{mk+r}}.$$

Moreover,

$$\beta_m^{(r)} = \alpha_{mk+r} \le C b^{mk+r} = C_1 b_1^m \quad (C_1 = C b^r,\ b_1 = b^k).$$

Hence, according to Lemma 1,

$$\frac{\alpha_{(mk+r)+k}}{\alpha_{mk+r}} = \frac{\beta_{m+1}^{(r)}}{\beta_m^{(r)}} \le a_r$$

for some $C_r, a_r \in (0, +\infty)$ and arbitrary m. Therefore,

$$\frac{\alpha_{n+k}}{\alpha_n} \le \max\{a_0, a_1, \ldots, a_{k-1}\} = a$$

for all n. The assertion is proved. □

Assertion 4. *Suppose that operators $B_1, \ldots, B_n$ act in X, $B_1, \ldots, B_{n-1}$ are linear operators with nonempty resolvent sets, $x \in X$, the series*

$$\sum_{m_1,\ldots,m_n=0}^{\infty} \frac{B_1^{m_1} \ldots B_n^{m_n}}{m_1! \ldots m_n!} x \tag{4}$$

converges absolutely, and the following condition holds:

(v) *for all $k \in \mathbb{N}$ and a set of natural indices $i_1, \ldots, i_k$ not exceeding n, the expression $B_{i_1} \ldots B_{i_k} x$ is valid and, if a set $j_1, \ldots, j_k$ is obtained from the set $i_1, \ldots, i_k$ by a rearrangement of its elements, then*

$$B_{i_1} \ldots B_{i_k} x = B_{j_1} \ldots B_{j_k} x. \tag{5}$$

In this case,

$$e^{B_1} \ldots e^{B_n} x = e^{B_1 + \cdots + B_n} x$$

(*both parts of the relation are valid*).

P r o o f. Let us first establish the equality

$$\sum_{m_1=0}^{\infty} \frac{B_1^{m_1}}{m_1!} \sum_{m_2=0}^{\infty} \frac{B_2^{m_2}}{m_2!} \cdots \sum_{m_n=0}^{\infty} \frac{B_n^{m_n}}{m_n!} x = \sum_{m_1,m_2,\ldots,m_n=0}^{\infty} \frac{B_1^{m_1} B_2^{m_2} \ldots B_n^{m_n}}{m_1! m_2! \ldots m_n!} x \tag{6}$$

by induction on n. For $n = 1$, (6) holds. Assume that, under the conditions of the assertion, equality (6) holds for $n = k - 1$ $(k \geq 2)$. Now, let $n = k$. Note that the absolute convergence of series (4) implies the absolute convergence of the series

$$\sum_{m_2,\ldots,m_n=0}^{\infty} \frac{B_2^{m_2} \ldots B_n^{m_n}}{m_2! \ldots m_n!} x.$$

Taking into account the fact that the operators $B_1^{m_1}$ $(m_1 \in \mathbb{N})$ are closed, condition (v), and the induction hypothesis, we obtain

$$\sum_{m_1=0}^{\infty} \frac{B_1^{m_1}}{m_1!} \sum_{m_2=0}^{\infty} \frac{B_2^{m_2}}{m_2!} \cdots \sum_{m_n=0}^{\infty} \frac{B_n^{m_n}}{m_n!} x = \sum_{m_1=0}^{\infty} \frac{B_1^{m_1}}{m_1!} \sum_{m_2,\ldots,m_n=0}^{\infty} \frac{B_2^{m_2} \ldots B_n^{m_n}}{m_2! \ldots m_n!} x =$$

$$= \sum_{m_1=0}^{\infty} \sum_{m_2,\ldots,m_n=0}^{\infty} \frac{B_1^{m_1} B_2^{m_2} \ldots B_n^{m_n}}{m_1! m_2! \ldots m_n!} x = \sum_{m_1,m_2,\ldots,m_n=0}^{\infty} \frac{B_1^{m_1} B_2^{m_2} \ldots B_n^{m_n}}{m_1! m_2! \ldots m_n!} x,$$

i. e., equality (6) is proved. Using this equality, we obtain

$$e^{B_1} \ldots e^{B_n} x = \sum_{m_1,\ldots,m_n=0}^{\infty} \frac{B_1^{m_1} \ldots B_n^{m_n}}{m_1! \ldots m_n!} x = \sum_{s=0}^{\infty} \sum_{m_1+\cdots+m_n=s} \frac{B_1^{m_1} \ldots B_n^{m_n}}{m_1! \ldots m_n!} x =$$

$$= \sum_{s=0}^{\infty} \frac{1}{s!} \sum_{\substack{m_1+\cdots+m_n=s \\ (m_1,\ldots,m_n \geq 0)}} \frac{s!}{m_1! \ldots m_n!} B_1^{m_1} \ldots B_n^{m_n} x = \sum_{s=0}^{\infty} \frac{(B_1 + \cdots + B_n)^s}{s!} x = e^{B_1+\cdots+B_n} x.$$

The assertion is proved. □

Remark 4. Equality (5) holds if the operators $B_1, \ldots, B_n$ pairwise commute and the left-hand side of (5) is valid.

Corollary 2. *Suppose that $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, $x \in X$, A is a linear operator acting in X, $\rho(A) \neq \varnothing$, and a series*

$$\sum_{m_1,\ldots,m_n=0}^{\infty} \frac{\alpha_1^{m_1} \ldots \alpha_n^{m_n}}{m_1! \ldots m_n!} A^{m_1+\cdots+m_n} x$$

converges absolutely. Then,

$$e^{\alpha_1 A} \ldots e^{\alpha_n A} x = e^{(\alpha_1+\cdots\alpha_n)A} x \tag{7}$$

(*both sides of the equality are valid*).

To formulate the next assertion, let us introduce some definitions and make a number of assumptions.

Suppose that $\mathscr{L} = \mathscr{L}(p, q)$ $(p > 0, q > 0)$ is a curve given in the complex plane (λ) by the equation

$$\beta^2 = 2p\alpha \ln \frac{\alpha}{q} \quad (\alpha = \operatorname{Re} \lambda,\ \beta = \operatorname{Im} \lambda,\ \alpha \geq q); \tag{8}$$

Let $G = G(p, q)$ be a domain containing the origin with the boundary $\mathscr{L}$; let the direction of $\mathscr{L}$ be chosen so that the domain G is on the right; and let A be an injective linear operator with domain $D(A)$ dense in X and range $\operatorname{Im}(A) \subset X$. The following estimate for the norm of the resolvent operator $R(\lambda) = R_A(\lambda) = (A - \lambda E)^{-1}$ of the operator A in $\overline{G}$ is known:

$$\|R(\lambda)\| \leq \frac{C_0}{(|\lambda| + 1)^{\gamma}} \tag{9}$$

for some $C_0 > 0$ and $\gamma \leq 1$ and all $\lambda \in \overline{G}$.

For $\varphi \in (0,\pi)$, we denote by $\Delta(\varphi)$ the domain in $\mathbb{C}$ that contains the negative real semiaxis and its boundary is $L(\varphi) = L_1(\varphi) \cup L_2(\varphi)$, where

$$L_1(\varphi) = \{\lambda \in \mathbb{C} : \lambda = te^{i\varphi}, t \geq 0\}, \quad L_2(\varphi) = \{\lambda \in \mathbb{C} : \lambda = te^{-i\varphi}, t \geq 0\}.$$

Suppose that $\Omega(a,\varphi) = \Delta(\varphi) \cup B(0,a)$ ($a > 0$, $\varphi \in (0,\pi)$, and $B(0,a)$ is the open disk of radius a centered at the origin), and the direction of $\Gamma(a,\varphi) = \partial\Omega(a,\varphi)$ is chosen so that $\Omega(a,\varphi)$ is on the right. Given p and q, we chose a and φ so that $\overline{\Omega(a,\varphi)} \subset G(p,q)$.

Under these assumptions, the authors studied [5] the operator functions

$$f(A) = -\frac{1}{2\pi i} A^n \int_{\Gamma(a,\varphi)} \frac{f(\lambda)}{\lambda^n} R(\lambda) d\lambda, \tag{10}$$

$$\widetilde{f}(A) = -\frac{1}{2\pi i} \overline{\int_{\Gamma(a,\varphi)} \frac{f(\lambda)}{\lambda^n} R(\lambda) d\lambda \; A^n} \tag{11}$$

constructed on the basis of corresponding scalar functions $f(\lambda)$ continuous in $\mathbb{C}\backslash\Omega(a,\varphi)$ and analytic in $\mathbb{C} \setminus \overline{\Omega(a,\varphi)}$; in addition, for every such function f there exist $C \in (0,+\infty)$ and $\sigma \in \mathbb{R}$ such that

$$|f(\lambda)| \leq C|\lambda|^\sigma \tag{12}$$

for all $\lambda \in \mathbb{C} \setminus \Omega(a,\varphi)$. The number $n \in \mathbb{N} \cup \{0\}$ in (10) and (11) is chosen so that $\sigma - n - \gamma < -1$.

It was proved that the right-hand sides of these representations are independent of such n, the operator functions $f(A)$ and $\widetilde{f}(A)$ are densely defined, $\widetilde{f}(A) \subset f(A)$, and the functions coincide if one of them is continuous.

We can take the function $e^{-t\lambda}$ ($t > 0$) as the function f and consider two operator functions, one of which is given by series according to formula (1) and the other is given by relations (10) and (11) for $n = 0$ (these relations yield the same result because their right-hand sides are continuous). Denoted by $(e^{-tA})_I$ the function given by formulas (10) and (11).

Lemma 2. *Let $\sigma \in \mathbb{R}$, $\sigma - \gamma < -1$, and let a function f be continuous in $\mathbb{C}\backslash\Omega(a,\varphi)$, analytic in $\mathbb{C}\backslash\overline{\Omega(a,\varphi)}$ and such that (12) holds for some $C \in (0,+\infty)$ and all $\lambda \in \mathbb{C}\backslash\Omega(a,\varphi)$.*

Then

$$\int_{\Gamma(a,\varphi)} f(\lambda)R(\lambda)d\lambda = \int_{\mathscr{L}(p,q)} f(\lambda)R(\lambda)d\lambda.$$

The proof of this lemma is similar to the proof of [3, Lemma 1].

Assertion 5. *Let a curve $\mathscr{L}_t$ ($t > 0$) be given by the equation $\beta^2 = 2tp\alpha \ln \dfrac{\alpha}{qt}$ ($\alpha = \operatorname{Re}\lambda$, $\beta = \operatorname{Im}\lambda$, and $\alpha \geq qt$), and let $n(t) \in \mathbb{N}$ satisfy the inequality*

$$tp - n(t) - \gamma < -1. \tag{13}$$

Then

$$e^{tA}(e^{-tA})_I\big|_{D(A^{n(t)})} = E\big|_{D(A^{n(t)})}, \tag{14}$$

$$(e^{-tA})_I e^{tA} = E\big|_{D(e^{tA})}. \tag{15}$$

P r o o f. Let us first consider the case $t = 1$. Note that $\mathscr{L}_1 = \mathscr{L}$. Denote by n_1^0 the value $n(1)$. The following equalities hold for $x \in D\left(e^A(e^{-A})_I\right)$:

$$e^A(e^{-A})_I x = -\frac{1}{2\pi i}\sum_{n=0}^{\infty}\frac{A^n}{n!}\int_{\mathscr{L}} e^{-\lambda}R(\lambda)d\lambda\ x = -\frac{1}{2\pi i}\lim_{n\to\infty}\sum_{k=0}^{n}\frac{A^k}{k!}\int_{\mathscr{L}} e^{-\lambda}R(\lambda)d\lambda\ x.$$

For every $n \in \mathbb{N}$, consider $n_1 \in \mathbb{N}$ satisfying the inequality $n - n_1 - \gamma < -1$. Then, according to [4, Theorem 9],

$$\sum_{k=0}^{n}\frac{A^k}{k!} = -\frac{A^{n_1}}{2\pi i}\int_{\mathscr{L}}\sum_{k=0}^{n}\frac{\lambda^{k-n_1}}{k!}R(\lambda)d\lambda;$$

i. e., according to [5, Theorem 3],

$$\sum_{k=0}^{n}\frac{A^k}{k!}\int_{\mathscr{L}} e^{-\lambda}R(\lambda)d\lambda = A^{n_1}\int_{\mathscr{L}}\sum_{k=0}^{n}\frac{\lambda^{k-n_1}}{k!}e^{-\lambda}R(\lambda)d\lambda. \tag{16}$$

For the functions $f_n(\lambda) = \sum_{k=0}^{n}\frac{\lambda^k}{k!}e^{-\lambda}$ in (10) and (11), we can take an arbitrary σ from (12). Thus, the right-hand side of (16) is independent of $n_1 \in \mathbb{N}\cup\{0\}$. Therefore, for $x \in D(A^{n_1^0})$, we have

$$A^{n_1}\int_{\mathscr{L}}\sum_{k=0}^{n}\frac{\lambda^{k-n_1}}{k!}e^{-\lambda}R(\lambda)d\lambda\ x = A^{n_1^0}\int_{\mathscr{L}}\sum_{k=0}^{n}\frac{\lambda^{k-n_1^0}}{k!}e^{-\lambda}R(\lambda)d\lambda\ x = \int_{\mathscr{L}}\frac{\lambda^{k-n_1^0}}{k!}e^{-\lambda}R(\lambda)d\lambda\ A^{n_1^0}x,$$

i. e.,

$$e^A(e^{-A})_I x = -\frac{1}{2\pi i}\lim_{n\to\infty}\int_{\mathscr{L}}\sum_{k=0}^{n}\frac{\lambda^{k-n_1^0}}{k!}e^{-\lambda}R(\lambda)d\lambda\ A^{n_1^0}x$$

whenever this limit exists. Let us establish the existence of this limit and find its value using the Lebesgue (dominated convergence) theorem on passing to the limit under the integral sign. Let us check the conditions of this theorem.

Let

$$H_n(\lambda) = \sum_{k=0}^{n}\frac{\lambda^{k-n_1^0}}{k!}e^{-\lambda}R(\lambda),$$

and let $\lambda = \alpha + i\beta \in \mathscr{L}$ be arbitrary. Then

$$\lim_{n\to\infty} H_n(\lambda) = \sum_{n=0}^{\infty}\frac{\lambda^{n-n_1^0}}{n!}e^{-\lambda}R(\lambda) = \lambda^{-n_1^0}R(\lambda)$$

(the limit and the convergence of the series are considered with respect to the operator norm). Let us show that the sequence $\{\|H_n(\lambda)\|\}$ is dominated by a Lebesgue integrable function in $\mathscr{L}$:

$$\|H_n(\lambda)\| \le \sum_{k=0}^{n}\frac{|\lambda|^{k-n_1^0}}{k!}e^{-\alpha}\|R(\lambda)\| \le \frac{C_0|\lambda|^{-n_1^0}e^{|\lambda|-\alpha}}{(|\lambda|+1)^{\gamma}} \le C_0C_1|\lambda|^{-n_1^0-\gamma}e^{|\lambda|-\alpha},$$

where $C_1 = \sup_{\mu\in\mathscr{L}}\left(\frac{|\mu|}{|\mu|+1}\right)^{\gamma}$ (the function $\left(\frac{|\mu|}{|\mu|+1}\right)^{\gamma}$ is continuous in $\mathscr{L}$, has a finite limit at infinity and, therefore, is bounded in $\mathscr{L}$). Using (8), we obtain that

$$|\lambda| - \alpha = \sqrt{\alpha^2+\beta^2} - \alpha = \frac{\beta^2}{\sqrt{\alpha^2+\beta^2}+\alpha} \le \frac{2p\alpha\ln\frac{\alpha}{q}}{2\alpha} = p\ln\frac{\alpha}{q};$$

i. e.,

$$e^{|\lambda|-\alpha} \le \left(\frac{\alpha}{q}\right)^p \le \left(\frac{|\lambda|}{q}\right)^p.$$

Hence,

$$\|H_n(\lambda)\| \le \overline{C}|\lambda|^{p-n_1^0-\gamma},$$

where $\overline{C} = \dfrac{C_0C_1}{q^p}$. By (13) for $t = 1$, the integral $\displaystyle\int_{\mathscr{L}} |\lambda|^{p-n_1^0-\gamma}|d\lambda|$ converges. By the Lebesgue theorem,

$$-\frac{1}{2\pi i}\lim_{n\to\infty}\int_{\mathscr{L}} H_n(\lambda)d\lambda = -\frac{1}{2\pi i}\int_{\mathscr{L}} \lambda^{-n_1^0}R(\lambda)d\lambda = A^{-n_1^0}$$

(the limit is considered with respect to the operator norm). Therefore,

$$e^A(e^{-A})_I x = A^{-n_1^0}A^{n_1^0}x = x,$$

and (14) is proved.

Let us show that (15) holds. For $x \in D(e^A) \subset \bigcap\limits_{n=0}^{\infty} D(A^n)$, in view of continuity of the operator $(e^{-A})_I$, we have

$$(e^{-A})_I e^A x = (e^{-A})_I \lim_{n\to\infty}\sum_{k=0}^{n}\frac{A^k x}{k!} = -\frac{1}{2\pi i}(e^{-A})_I\int_{\mathscr{L}}\sum_{k=0}^{n}\frac{\lambda^{k-n_1}}{k!}R(\lambda)d\lambda\ A^{n_1}x,$$

where n_1, as before, satisfies inequality (13) for $t = 1$. Arguing similarly to the proof of formula (14), we obtain (15). Thus, the assertion holds for $t = 1$.

Let us now consider an arbitrary $t > 0$. The mapping $\mu = t\lambda$ takes the curve $\mathscr{L}$ to the curve $\mathscr{L}_t$ and the domain $G = G_1$ to the domain $G_t \ni 0$ such that $\partial G_t = \mathscr{L}_t$. In addition, $\rho(tA) = t\rho(A)$ ($\rho(A)$ and $\rho(tA)$ are the regular sets of the operators A and tA, respectively) and the estimate for $\|R_{tA}(\lambda)\|$ in G_t coincides with the estimate (9) for $\|R_A(\lambda)\|$ in G with certain constant C_t instead of C_0. The analysis of the proof for $t = 1$ shows that formulas (14) and (15) remain valid for $t > 0$ under condition (13). The assertion is proved. □

Corollary 3. *If $t > 0$ and the operator e^{tA} is closed, then it is invertible and $(e^{tA})^{-1} = (e^{-tA})_I$.*

The corollary follows from (14), (15), and the fact that if a closed operator coincides with a continuous operator on a dense set, then they coincide in the entire space.

Example 2. Let $X = L_p[1,+\infty)$ and $Ax(t) = tx(t)$ $(x \in X)$. Let us show that $\mathcal{D}(e^A) = \{x \in X : e^t x \in X\}$ and the equality $e^A x = e^t x$ holds for $x \in \mathcal{D}(e^A)$.

Let $x \in X$ and$e^t x \in X$. Let us establish that $x \in \mathcal{D}(e^A)$ and $e^A x = e^t x$. To this end, we have to prove that the series $\displaystyle\sum_{n=0}^{\infty}\frac{t^n}{n!}x$ converges in X to $e^t x$, i. e., that

$$\left\|e^t x - \sum_{k=0}^{n}\frac{A^k x}{k!}\right\| = \left\|e^t x - \sum_{k=0}^{n}\frac{t^k x}{k!}\right\| \xrightarrow[k\to\infty]{} 0.$$

Since $\displaystyle\int_1^{+\infty} e^{pt}|x(t)|^p\,dt < +\infty$, there exists a function $\alpha(t)$ defined on $[1,+\infty)$ such that $\alpha(t) \ge \alpha_0$ for some $\alpha_0 > 0$, $\alpha(t) \xrightarrow[t\to\infty]{} +\infty$ (in particular, we can take a continuous positive function α with infinite limit at $+\infty$), and $\displaystyle\int_1^{+\infty} |\alpha(t)e^t x(t)|^p\,dt < +\infty$. Then

$$\left\|e^t x - \sum_{k=0}^{n}\frac{t^k x}{k!}\right\| = \left\|\frac{1 - e^{-t}\sum_{k=0}^{n} t^k/k!}{\alpha}\alpha e^t x\right\| \le \sup_{t\ge 1}\frac{1 - e^{-t}\sum_{k=0}^{n} t^k/k!}{\alpha(t)}\|\alpha e^t x\| = \gamma_n\|\alpha e^t x\|.$$

Let us show that

$$\gamma_n = \sup_{t\geq 1} \frac{1-e^{-t}\sum_{k=0}^{n} t^k/k!}{\alpha(t)} \xrightarrow[n\to\infty]{} 0.$$

Take an arbitrary $\varepsilon > 0$. Since $\alpha(t) \xrightarrow[t\to\infty]{} +\infty$, there is a number $\Delta > 1$ such that $1/\alpha(t) < \varepsilon$ for all $t \geq \Delta$; i. e.,

$$\frac{1-e^{-t}\sum_{k=0}^{n} t^k/k!}{\alpha(t)} \in [0,\varepsilon)$$

for all $t \geq \Delta$ and $n \in \mathbb{N}$. Since the power series $\sum_{n=0}^{\infty} t^n/n!$ uniformly converges to e^t on $[1,\Delta]$, the sequence of functions

$$\left\{\frac{1-e^{-t}\sum_{k=0}^{n} t^k/k!}{\alpha(t)}\right\}$$

uniformly tends to zero on $[1,\Delta]$. Hence, there exists a number N such that

$$\frac{1-e^{-t}\sum_{k=0}^{n} t^k/k!}{\alpha(t)} < \varepsilon$$

for all $t \in [1,\Delta]$ and $n > N$. Thus, $\gamma_n \leq \varepsilon$ for all $n > N$, i. e., $\gamma_n \xrightarrow[n\to\infty]{} 0$ and, consequently,

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} x = e^t x \quad \text{in} \quad X.$$

Conversely, suppose that $x \in \mathcal{D}(e^A)$ and $e^A x = y \in X$, i. e.,

$$S_n = \sum_{k=0}^{n} \frac{t^k}{k!} x \xrightarrow[n\to\infty]{} y \quad \text{in} \quad X.$$

Then, there exists a subsequence $\{S_{n_k}\}$ of $\{S_n\}$ such that

$$S_{n_k}(t) \xrightarrow[k\to\infty]{\text{a. e.}} y(t).$$

But

$$S_n(t)x(t) \xrightarrow[n\to\infty]{} e^t x(t)$$

at every point $t \geq 1$; i. e., $y = e^t x$ in X. Since $e^t x \in X$, we have $e^A x = e^t x$.

Note that equality (7) holds for the operator A if

$$x,\ e^{\alpha_n t}x,\ e^{(\alpha_{n-1}+\alpha_n)t}x, \ldots, e^{(\alpha_1+\cdots+\alpha_n)t}x \in X.$$

Conclusion

We have considered some natural properties of exponential operator defined by power series (Corollary 1 and Assertion 4). The main result of the paper is the connection (under certain conditions) of the exponential operator e^A in the form of power series with the exponential function e^{-A} defined on the basis of the Cauchy integral formula (Assertion 5). These facts may give an impulse to obtaining further results on functional calculus of operators.

REFERENCES

1. Balakrishnan A. V. Fractional powers of closed operators and the semigroups generated by them. *Pacific J. Math.*, 1960. Vol. 10, No. 2. P. 419–437. URL: https://projecteuclid.org/euclid.pjm/1103038401
2. Dunford N., Schwartz J. T. *Linear Operators Part I: General Theory.* New York: Interscience Publishers, 1958. 858 p.
3. Korkina L. F., Rekant M. A. An extension of the class of power operator functions. *Izvestiya Uralskogo gosudarstvennogo universiteta* (*Matematika i mekhanika*) [Bulletin of the Ural State University (Mathematics and Mechanics)], 2005. No. 38. P. 80–90. (in Russian) URL: http://hdl.handle.net/10995/24591
4. Korkina L. F., Rekant M. A. Some classes of functions of a linear closed operator. *Proc. Steklov Inst. Math.*, 2012. Vol. 277, Suppl. 1. P. 121–135. DOI: 10.1134/S0081543812050124.
5. Korkina L. F., Rekant M. A. Properties of mappings of scalar functions to operator functions of a linear closed operator. *Trudy Inst. Mat. i Mekh. UrO RAN* [Proc. of Krasovskii Institute of Mathematics and Mechanics of the UB RAS], 2015. Vol. 21, No. 1. P. 153–165. (in Russian) URL: http://mi.mathnet.ru/eng/timm/v21/i1/p153
6. Krein S. G. *Lineinye differentsial'nye uravneniya v banakhovom prostranstve* [Linear Differential Equations in Banach Space]. Moscow: Nauka, 1967. 464 p. (in Russian)
7. Lusternik L. A., Sobolev V. J. *Elements of functional analysis.* Delhi: Hindustan Publishing Corpn., 1974. 376 p.
8. Rudin W. *Functional Analysis.* New York: McGraw-Hill, 1973. 397 p.

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 43–55
DOI: 10.15826/umj.2018.2.006

OPTIMIZING THE STARTING POINT IN A PRECEDENCE CONSTRAINED ROUTING PROBLEM WITH COMPLICATED TRAVEL COST FUNCTIONS[1]

Alexander G. Chentsov†, Alexey M. Grigoryev†† and Alexey A. Chentsov†††

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990

†chentsov@imm.uran.ru, ††ag@uran.ru, †††chentsov_a_a@mail.ru

Abstract: We study the optimization of the initial state, route (a permutation of indices), and track in an extremal problem connected with visiting a finite system of megalopolises subject to precedence constraints where the travel cost functions may depend on the set of (pending) tasks. This problem statement is exemplified by the task to dismantle a system of radiating elements in case of emergency, such as the Chernobyl or Fukushima nuclear disasters. We propose a solution based on a parallel algorithm, which was implemented on the Uran supercomputer. It consists of a two-stage procedure: stage one determines the value (extremum) function over the set of all possible initial states and finds its minimum and also the point where it is achieved. This point is viewed as a base of the optimal process, which is constructed at stage two. Thus, optimization of the starting point for the route through megalopolises, connected with conducting certain internal tasks there, is an important element of the solution. To this end, we employ the apparatus of the broadly understood dynamic programming with elements of parallel structure during the construction of Bellman function layers.

Key words: Dynamic programming, Route, Sequencing, Precedence constraints, Parallel computation.

Introduction

In this paper, we consider an additive routing problem aimed at applications in nuclear power generation: the intention is to decrease the exposure of power plant staff to radiation during a sequence of work-related activities. The considered problem features precedence constraints, multiple variants of movements, and travel cost functions that could depend on the set of pending tasks. The mentioned features of the statement stem from the peculiarities of the actual engineering problem, which exhibits a qualitative difference from its prototype, the well-known intractable traveling salesman problem (TSP); see [1–6]. In a series of papers, the authors have developed solution methods based on dynamic programming (DP) combined with parallel computations, see [7–12] et al. Here, we consider the statement where, in addition to a solution in the form of a route-track pair, one also has to choose the starting point (the base) for the process of movements. We found out that the DP-based procedure used in [7–11] could be used just as well to solve such an "expanded" problem (see also the monograph [13], connected with issues of decreasing staff exposure to radiation during a sequence of operations).

1. General notation and definitions

We use the quantifiers and logical connectives; $\emptyset$ denotes the empty set and $\triangleq$ denotes the equality by definition. To arbitrary objects α and β, assign the set $\{\alpha;\beta\}$ that contains α and β and them only. If x is an object, then $\{x\} \triangleq \{x;x\}$ is the singleton that contains x. A set is an object,

[1]This work was supported by Russian Science Foundation (project no. 14-11-00109).

hence [14, p. 59], to objects y and z, one can assign an ordered pair (OP) $(y, z) \triangleq \{\{y\}; \{y; z\}\}$ of these objects; y is the first and z is the second element of this OP. To every OP z, assign the first element $\mathrm{pr}_1(z)$ and the second element $\mathrm{pr}_2(z)$, which are uniquely defined by the condition $z = \big(\mathrm{pr}_1(z), \mathrm{pr}_2(z)\big)$. If a, b, and c are objects, then $(a, b, c) \triangleq \big((a, b), c\big)$ is the triple of these objects, constructed as an OP with the first element (a, b) and the second element c.

To every set S, assign the family $\mathcal{P}(S)$ of all its subsets; $\mathcal{P}'(S) \triangleq \mathcal{P}(S) \setminus \{\emptyset\}$; and $\mathrm{Fin}(S)$ is the family of all finite sets from $\mathcal{P}'(S)$. The family $\mathrm{Fin}(S)$ consists of the finite nonempty subsets of S and them only. To nonempty sets A and B, assign the nonempty set B^A of all mappings from A to B (see [14, p. 70]); for $g \in B^A$ and $C \in \mathcal{P}(A)$, in the form $g^1(C) \triangleq \{g(x) : x \in C\} \in \mathcal{P}(B)$, we have the image of C under g; $g^1(C) \neq \emptyset$ when $C \neq \emptyset$. If A, B, and C are three nonempty sets, then [15, p. 5] $A \times B \times C \triangleq (A \times B) \times C$; if, in addition, D is a nonempty set and $h \in D^{A \times B \times C}$, then, for $x \in A \times B$ and $y \in C$, we have $(x, y) \in A \times B \times C$ and the value $h(x, y) \in D$ of the mapping h at the point (x, y) is well-defined; for this value, we also have $h(x, y) = h\big(\mathrm{pr}_1(x), \mathrm{pr}_2(x), y\big)$.

As usual, $\mathbb{N} \triangleq \{1; 2; \ldots\}$; set $\mathbb{N}_0 \triangleq \{0\} \cup \mathbb{N}$ and $\overline{p, q} \triangleq \{j \in \mathbb{N}_0 | \, (p \leqslant j) \& (j \leqslant q)\}$ $\forall p \in \mathbb{N}_0$ $\forall q \in \mathbb{N}_0$ (if $k \in \mathbb{N}_0$, $l \in \mathbb{N}_0$, and $l < k$, then $\overline{k, l} = \emptyset$). To every nonempty finite set K assign its cardinality $|K| \in \mathbb{N}$ and the nonempty set $(\mathrm{bi})[K]$ of all bijections of the integer interval $\overline{1, |K|}$ onto K; $|\emptyset| \triangleq 0$. By $\mathbb{R}$ we denote the real line; $\mathbb{R}_+ \triangleq \{\xi \in \mathbb{R} | \, 0 \leqslant \xi\}$; and $\mathcal{R}_+[T]$, where T is a nonempty set, denotes the set of all functions from T to $\mathbb{R}_+$, that is, $\mathcal{R}_+[T] \triangleq (\mathbb{R}_+)^T$.

2. Problem statement

Fix a nonempty set X and some $X^0 \in \mathrm{Fin}(X)$; the points X^0 are viewed as admissible starting points. Let $N \in \mathbb{N}$, $N \geqslant 2$,

$$M_1 \in \mathrm{Fin}(X), \ldots, M_N \in \mathrm{Fin}(X),$$

and let $\mathbb{M}_1 \in \mathcal{P}'(M_1 \times M_1), \ldots, \mathbb{M}_N \in \mathcal{P}'(M_N \times M_N)$; assume

$$(X^0 \cap M_j = \emptyset \quad \forall j \in \overline{1, N}) \& (M_p \cap M_q = \emptyset \quad \forall p \in \overline{1, N} \quad \forall q \in \overline{1, N} \setminus \{p\})$$

and set $\mathbb{P} \triangleq (\mathrm{bi})[\overline{1, N}]$. Consider the processes

$$\begin{gathered}(x^{(0)} = x^0 \in X^0) \to (x_1^{(1)} \in M_{\alpha(1)} \rightsquigarrow x_2^{(1)} \in M_{\alpha(1)}) \to \ldots \\ \to (x_1^{(N)} \in M_{\alpha(N)} \rightsquigarrow x_2^{(N)} \in M_{\alpha(N)})\end{gathered} \tag{2.1}$$

where $\alpha \in \mathbb{P}$, $z_1 \in (x_1^{(1)}, x_2^{(1)}) \in \mathbb{M}_{\alpha(1)}, \ldots, z_N \in (x_1^{(N)}, x_2^{(N)}) \in \mathbb{M}_{\alpha(N)}$. The permutation α determines the route, that is, the sequence the megalopolises are visited in while $z_1, \ldots, z_N$ determines the track of these visits; x^0 is the initial state. A complete solution (see (2.1)) is a tuple $(x^0, \alpha, z_1, \ldots, z_N)$, to be determined. The choice of $\alpha \in \mathbb{P}$ may be restricted by precedence constraints; to describe them, fix

$$\mathbf{K} \in \mathcal{P}(\overline{1, N} \times \overline{1, N}),$$

that is, the set of OPs known as "address pairs" (see [7–11, 13]); the case $\mathbf{K} = \emptyset$ denotes the absence of precedence constraints. Assume that

$$\forall \mathbf{K}_0 \in \mathcal{P}'(\mathbf{K}) \; \exists z_0 \in \mathbf{K}_0 : \; \mathrm{pr}_1(z_0) \neq \mathrm{pr}_2(z) \quad \forall z \in \mathbf{K}_0.$$

In the form

$$\mathbf{A} \triangleq \{\alpha \in \mathbb{P} | \, \alpha^{-1}\big(\mathrm{pr}_1(z)\big) < \alpha^{-1}\big(\mathrm{pr}_2(z)\big) \quad \forall z \in \mathbf{K}\} \in \mathcal{P}'(\mathbb{P}), \tag{2.2}$$

we have a (nonempty) set of $\mathbf{K}$-feasible routes. Let $\mathbb{X} \triangleq X^0 \cup (\bigcup_{i=1}^{N} M_i)$; then, $\mathbb{X} \in \mathrm{Fin}(X)$. Denote by $\mathbb{Z}$ the set of all tuples $(z_i)_{i\in\overline{0,N}} : \overline{0,N} \to \mathbb{X}\times\mathbb{X}$, that is, $\mathbb{Z} \triangleq (\mathbb{X}\times\mathbb{X})^{\overline{0,N}}$. If $x^0 \in X^0$ and $\alpha \in \mathbb{P}$, then

$$\mathcal{Z}_\alpha[x^0] \triangleq \{\mathbf{z} \in \mathbb{Z} |\, \big(\mathbf{z}(0) = (x^0, x^0)\big) \& \big(\mathbf{z}(t) \in \mathbb{M}_{\alpha(t)} \quad \forall t \in \overline{1,N}\big)\} \in \mathrm{Fin}(\mathbb{Z}). \tag{2.3}$$

Therefore, for $x^0 \in X^0$, in the form

$$\widetilde{\mathbf{D}}[x^0] \triangleq \{(\alpha, \mathbf{z}) \in \mathbf{A}\times\mathbb{Z} |\, \mathbf{z} \in \mathcal{Z}_\alpha[x^0]\} \in \mathrm{Fin}(\mathbf{A}\times\mathbb{Z}),$$

we have a (nonempty finite) set of feasible solutions (FS) of the problem with the fixed initial state. Next, let us note that

$$\mathbf{D} \triangleq \{(\alpha, \mathbf{z}, x) \in \mathbf{A}\times\mathbb{Z}\times X^0 |\, (\alpha, \mathbf{z}) \in \widetilde{\mathbf{D}}[x]\} \in \mathrm{Fin}(\mathbf{A}\times\mathbb{Z}\times X^0) \tag{2.4}$$

is viewed as the set of all FSs of the complete problem.

Consider the following transportation cost functions. Let $\mathfrak{N} \triangleq \mathcal{P}'(\overline{1,N})$; $\mathbf{c} \in \mathcal{R}_+[\mathbb{X}\times\mathbb{X}\times\mathfrak{N}]$; and let $c_1 \in \mathcal{R}_+[\mathbb{X}\times\mathbb{X}\times\mathfrak{N}], \ldots, c_N \in \mathcal{R}_+[\mathbb{X}\times\mathbb{X}\times\mathfrak{N}]$, $f \in \mathcal{R}_+[\mathbb{X}]$. In terms of the tuple

$$(\mathbf{c}, c_1, \ldots, c_N, f),$$

we define the additive criterion: for $x^0 \in X^0$ and $(\alpha, \mathbf{z}) \in \widetilde{\mathbf{D}}[x^0]$, assume

$$\begin{aligned}\mathfrak{C}_\alpha[\mathbf{z}] \triangleq \sum_{s=1}^{N} \Big[\mathbf{c}\Big(\mathrm{pr}_2\big(\mathbf{z}(s-1)\big), \mathrm{pr}_1\big(\mathbf{z}(s)\big), \alpha^1(\overline{s,N})\Big) + \\ + c_{\alpha(s)}\big(\mathbf{z}(s), \alpha^1(\overline{s,N})\big)\Big] + f\Big(\mathrm{pr}_2\big(\mathbf{z}(N)\big)\Big);\end{aligned} \tag{2.5}$$

thus, to each FS $(\alpha, \mathbf{z}, x^0) \in \mathbf{D}$, we assign the value $\mathfrak{C}_\alpha[\mathbf{z}] \in \mathbb{R}_+$, which does not explicitly depend on x^0 (x^0 affects the choice of $\mathbf{z}$). Like in [7–11], for $x^0 \in X^0$, let us introduce the problem

$$\mathfrak{C}_\alpha[\mathbf{z}] \longrightarrow \min, \quad (\alpha, \mathbf{z}) \in \widetilde{\mathbf{D}}[x^0], \tag{2.6}$$

for which the value $V[x^0]$ is determined as the least value among $\mathfrak{C}_\alpha[\mathbf{z}]$, $(\alpha, \mathbf{z}) \in \widetilde{\mathbf{D}}[x^0]$, and also the (nonempty) set

$$(\mathrm{SOL})[x^0] \triangleq \{(\alpha_0, \mathbf{z}_0) \in \widetilde{\mathbf{D}}[x^0] |\, \mathfrak{C}_{\alpha_0}[\mathbf{z}_0] = V[x_0]\} \in \mathrm{Fin}(\widetilde{\mathbf{D}}[x^0]). \tag{2.7}$$

In addition, we have the following complete problem

$$\mathfrak{C}_\alpha[\mathbf{z}] \longrightarrow \min \quad (\alpha, \mathbf{z}, x) \in \mathbf{D}, \tag{2.8}$$

with the value

$$\mathbb{V} \triangleq \min_{(\alpha,\mathbf{z},x)\in\mathbf{D}} \mathfrak{C}_\alpha[\mathbf{z}] \in \mathbb{R}_+ \tag{2.9}$$

and a (nonempty) set

$$\mathbb{SOL} \triangleq \{(\alpha^0, \mathbf{z}^0, x^0) \in \mathbf{D} |\, \mathfrak{C}_{\alpha^0}[\mathbf{z}^0] = \mathbb{V}\} \in \mathrm{Fin}(\mathbf{D}).$$

In connection with (2.8), it is also of interest to consider the problem

$$V[x] \longrightarrow \min, \quad x \in X^0; \tag{2.10}$$

(2.10) is the problem of optimizing the starting point, which is of some interest in itself. Indeed, if (2.10) is solved, we get $\mathbb{V}$ (2.9) and the point $x^0 \in X^0$ such that $\mathbb{V}$ is achieved by $V[x^0]$. One could construct heuristics (when necessitated by the problem's dimension) for solving (2.6), compare their results with $\mathbb{V}$, and thereby choose what is deemed admissible. In this connection, note that (see (2.4))

$$\mathbb{V} = \min_{x \in X^0} \min_{(\alpha, \mathbf{z}) \in \widetilde{\mathbf{D}}[x]} \mathfrak{C}_\alpha[\mathbf{z}] = \min_{x \in X^0} V[x]. \tag{2.11}$$

To solve the problems of the form (2.6), one can use the broadly understood DP in the spirit of [7–11, 13]; here, we consider these procedures in their algorithmic form (see [11, 16]).

3. Dynamic programming in starting point optimization problem

This section serves to adapt the DP procedure from papers [7–11, 13, 16] to the needs of solving problem (2.10). To this end, let us introduce the crossing-out operator $\mathbf{I}$, which acts in $\mathfrak{N}$: for $K \in \mathfrak{N}$, assume

$$\mathbf{I}(K) \triangleq K \setminus \{\mathrm{pr}_2(z) : z \in \Xi[K]\}, \tag{3.1}$$

where $\Xi[K] \triangleq \{z \in \mathbf{K} | \,(\mathrm{pr}_1(z) \in K)\&(\mathrm{pr}_2(z) \in K)\}$ (note that $\mathbf{I}(\{t\}) = \{t\}$ for $t \in \overline{1,N}$). In terms of $\mathbf{I}$ (3.1), let us introduce the family

$$\mathfrak{C} \triangleq \{K \in \mathfrak{N} | \,\forall z \in \mathbf{K} \ \ (\mathrm{pr}_1(z) \in K) \Rightarrow (\mathrm{pr}_2(z) \in K)\}$$

of feasible (task) sets and its subfamilies $\mathfrak{C}_s \triangleq \{K \in \mathfrak{C} | \, s = |K|\} \quad \forall s \in \overline{1,N}$. Note that $\mathfrak{C}_N = \{\overline{1,N}\}$ and $\mathfrak{C}_{s-1} = \{K \setminus \{t\} : K \in \mathfrak{C}_s, \ t \in \mathbf{I}(K)\} \quad \forall s \in \overline{2,N}$ (we have (see [16]) a recurrence procedure for constructing $\mathfrak{C}_1, \ldots, \mathfrak{C}_N$). For $\mathbf{K}_1 \triangleq \{\mathrm{pr}_1(z) : z \in \mathbf{K}\}$, we have the equality $\mathfrak{C}_1 = \big\{\{t\} : t \in \overline{1,N} \setminus \mathbf{K}_1\big\}$. Let $\mathbf{M}_t \triangleq \{\mathrm{pr}_2(z) : z \in \mathbb{M}_t\} \ \ \forall t \in \overline{1,N}$. In addition, let

$$\mathbf{X} \triangleq X^0 \cup \Big(\bigcup_{t=1}^{N} \mathbf{M}_t\Big). \tag{3.2}$$

Consider the construction of layers of the state space, that is, the layers of the set $\mathbf{X} \times \mathcal{P}(\overline{1,N})$. To this end, first, denote by $\widetilde{\mathcal{M}}$ the union of all the sets $\mathbf{M}_t$, $t \in \overline{1,N} \setminus \mathbf{K}_1$; then, set

$$D_0 \triangleq \{(x, \emptyset) : x \in \widetilde{\mathcal{M}}\}.$$

In addition, set $D_N \triangleq \{(x, \overline{1,N}) : x \in X^0\}$; D_0 and D_N are the boundary state space layers.

Constructing intermediary layers. If $s \in \overline{1,N-1}$ and $K \in \mathfrak{C}_s$, then let us define, in a sequential fashion, the three sets

$$\begin{gathered} J_s(K) \triangleq \{j \in \overline{1,N} \setminus K | \,\{j\} \cup K \in \mathfrak{C}_{s+1}\}, \\ \mathcal{M}_s[K] \triangleq \bigcup_{j \in J_s(K)} \mathbf{M}_j, \\ \mathbb{D}_s[K] \triangleq \{(x, K) : x \in \mathcal{M}_s[K]\}. \end{gathered} \tag{3.3}$$

In view of (3.3), for $s \in \overline{1,N-1}$, let D_s be the union of all the sets $\mathbb{D}_s[K]$, $K \in \mathfrak{C}_s$; then, $\emptyset \neq D_s \subset \mathbf{X} \times \mathfrak{C}_s$.

In view of the definitions of D_0 and D_N, we see that, in particular, $(D_s)_{s\in\overline{0,N}}$ is a tuple of subsets of $\mathbf{X}\times\mathcal{P}(\overline{1,N})$. Thus we obtain the state space layers. Let us now define the functions

$$v_0\in\mathcal{R}_+[D_0], v_1\in\mathcal{R}_+[D_1],\dots,v_N\in\mathcal{R}_+[D_N]$$

in a sequential fashion. Set $v_0(x,\emptyset)\overset{\triangle}{=}f(x)\ \forall x\in\widetilde{\mathcal{M}}$; thus, we obtain v_0.

Under $s\in\overline{1,N}$, $(x,K)\in D_s$, $j\in\mathbf{I}(K)$, and $z\in\mathbb{M}_j$, we obtain (see [16, (4.9)])

$$\big(\mathrm{pr}_2(z),K\setminus\{j\}\big)\in D_{s-1}.$$

In view of this property, for $s\in\overline{1,N}$, we define the transformation of v_{s-1} into v_s through [16, Proposition 4.1]:

$$\begin{aligned} v_s(x,K)\overset{\triangle}{=}\min_{j\in\mathbf{I}(K)}\ \min_{z\in\mathbb{M}_j}\big[\mathbf{c}\big(x,\mathrm{pr}_1(z),K\big)+c_j(z,K)+\\ +v_{s-1}\big(\mathrm{pr}_2(z),K\setminus\{j\}\big)\big]\quad\forall(x,K)\in D_s.\end{aligned}\tag{3.4}$$

This implements the recurrence procedure $v_0\to v_1\to\ldots\to v_N$.

Proposition 3.1. *If $x^0\in X^0$, then $v_N(x^0,\overline{1,N})=V[x^0]$.*

P r o o f. Fix $x^0\in X^0$, which implies $x^0\in X$. Consider problem (2.6). The way of solving this problem is described, in particular, in [16]; in the same paper, there are also constructed the feasible task set families $\mathfrak{C},\mathfrak{C}_1,\dots,\mathfrak{C}_N$ similar to those mentioned in the beginning of the section. Based on that (in [16]), state space layers $\mathcal{D}_0,\mathcal{D}_1,\dots,\mathcal{D}_N$, similar to $D_0,D_1,\dots,D_N$ (see, in particular, (3.3) and [16, Section 4]), are constructed. There is a difference only for D_N and $\mathcal{D}_N$: here, we have

$$D_N=X^0\times\{\overline{1,N}\}=\{(x,\overline{1,N}):x\in X^0\},$$

whereas, in [16], $\mathcal{D}_N=\{(x^0,\overline{1,N})\}$, whence $\mathcal{D}_N\subset D_N$. Next, the construction of $v_0,v_1,\dots,v_{N-1}$ in [16, Section 4] and in the present section is the same (see, in particular, (3.4) and [16, Proposition 4.1]). Therefore, in particular, v_{N-1} matches that of [16, Section 4]. At the same time, $V[x^0]$ matches V [16, (3.18)]. Thus, in accordance with [16, (4.12)],

$$V[x^0]=\min_{j\in\mathbf{I}(\overline{1,N})}\ \min_{z\in\mathbb{M}_j}\big[\mathbf{c}\big(x^0,\mathrm{pr}_1(z),\overline{1,N}\big)+c_j(z,\overline{1,N})+v_{N-1}\big(\mathrm{pr}_2(z),\overline{1,N}\setminus\{j\}\big)\big],\tag{3.5}$$

where $\big(\mathrm{pr}_2(\widetilde{z}),\overline{1,N}\setminus\{j\}\big)\in D_{N-1}$ for $j\in\mathbf{I}(\overline{1,N})$ and $\widetilde{z}\in\mathbb{M}_j$. However, $(x^0,\overline{1,N})\in D_N$, thus, in the right-hand side of (3.5), we have (see (3.4)) $v_N(x^0,\overline{1,N})$, which completes the proof. □

From Proposition 3.1, we see that problem (2.10) takes the following form:

$$v_N(x,\overline{1,N})\longrightarrow\min,\quad x\in X^0.\tag{3.6}$$

In (3.6), we have an exhaustive search for the minimum of the function $v_N(\cdot,\overline{1,N})$ over the finite set X^0. In this connection, we propose the following algorithm for solving problem (2.10).

4. Algorithm for optimization of starting point

Algorithm 4.1.

(1) In terms of f, define the function v_0.

(2) If $s \in \overline{1, N}$ and the function v_{s-1} have been constructed already, conduct the transformation $v_{s-1} \to v_s$ based on (3.4).

(3) After v_s has been constructed through the rule (3.4), the values of the function v_{s-1} are erased and replaced by the values of the function v_s (the Bellman function layers are overwritten).

(4) After the function v_N has been constructed, solve the problem (3.6): determine $\mathbb{V}$ and the minimum of the function $v_N(\cdot, \overline{1, N})$, which has the form

$$x \longmapsto v_N(x, \overline{1, N}) : X^0 \to \mathbb{R}_+.$$

As noted before, the solution of problem (2.10) could be used to test the heuristics, which are to be employed on larger problem instances.

Coming back to the problem (2.8), note that the aforementioned algorithm (which admits a natural analogy with [16]) must be modified: the layers $v_1, \dots, v_N$ will have to be retained in the computer's memory. We also have to select the point $x^0 \in X^0$ that is a solution of problem (2.10), that is, $V[x^0] = \mathbb{V}$. Next, use the algorithm [16, Section 4] (see also [17, § 7], where a slightly more general statement was considered). The logic of constructions here follows that of [7–11, 13].

To construct an optimal solution after the optimal starting point has been found — the pair of a route and a track — we use the algorithm [16, Section 4] (see also [7, 11]). In this case, we have to retain in the computer's memory all the layers of the corresponding (to the found starting point) "part" of the Bellman function. At this stage, it is also possible to repeat the construction of the layers of the mentioned "part" that corresponds to the solution of problem (3.6). We omit this construction and refer the reader to [7, 16, 19] for details.

Using the independent computations scheme. Returning to problem (3.6), note that its most significant step — the construction of the Bellman function layers — is conducted through the independent computations scheme (see papers [17, 18]), which transfers to problem (3.6) without significant modifications because the actual object of construction in [17, 18] (and also [9, 11]) is the function v_{N-1}. Thus, we omit the theoretical description of the independent computations scheme in the spirit of [17, 18], and the parallel algorithm itself is only briefly described in connection with its software implementation for a supercomputer. The differences with [17, 18] only appear in the final computations of the form (3.5) (in [17, 18], a single computation was required, whereas, for problem (3.6), the number of computations matches the number of elements in the set X^0). A version of parallel implementation as described in [17, 18] can be used both for solving problem (2.10), (3.6) (when the Bellman function layers get overwritten, see step (1)–(4) of the Algorithm 4.1), and in subsequent construction of the optimal solution in the form of a route-track pair "tied" to the minimum of problem (3.6). Following [17, 18], we distribute the sets from $\mathfrak{C}_{N-1}$ between the nodes, creating thus a finite collection of independent computation procedures; these procedures could, in part, overlap (the layers $D_s, s \in \overline{1, N-1}$ are covered by the "individual" state space layers, which do not normally reduce to a partition; the systems of individual layers, each connected with a fixed set $K \in \mathfrak{C}_{N-1}$, denote the "theaters" of the corresponding nodes). Each of the mentioned computational procedures yields a "part" (to be more precise, a restriction to a nonempty subset of D_{N-1}) of the function v_{N-1}.

Application to a dismantling problem. One natural version of the general statement can be connected with the problem of dismantling, one by one, in a sequence, a finite system of radiating elements. The goal is to minimize the total radiation dose incurred by a staff member, by means of selecting the starting point, the route (in the form of a permutation of indices), and the actual trajectory. In this special case, problem (2.10), (3.6) has the following sense: namely, where specifically should the agent (or a crew) be brought to minimize the total radiation dose in view of the subsequent optimization of the route and track. At the same time, technology-determined precedence constraints have to be satisfied, and the travel costs present a rather complicated form of dependence on the set of tasks that have not been completed yet at the time of travel. Let us now discuss one fragment of the construction of the mentioned function.

Cost function. Assume that, on a plane, there are given the points x and y, $x \neq y$; consider the travel from x to y assuming the set $K \in \mathrm{Fin}(X)$, where $X = \mathbb{R} \times \mathbb{R}$, is formed by the radiation sources that are have not been dismantled yet. Then, the radiation dose $\mathbf{c}(x, y, K) \in \mathbb{R}_+$ for the mentioned motion is obtained by summing the values $\mathbf{c}(x, y, \{z\}) \in \mathbb{R}_+$ for $z \in K$. Each single value $\mathbf{c}(x, y, \{u\})$, for some $u \in K$, has, in the "regular" case, the following form:

$$\mathbf{c}(x, y, \{u\}) = \gamma_u \int_0^T \frac{1}{(\rho(u, w_t))^2} dt. \tag{4.1}$$

Here ρ denotes the Euclidean distance in X; $\gamma_u \in \mathbb{R}_+, \gamma_u \neq 0$, is the coefficient that determines the intensity of the source $u \in K$; the travel time T is uniquely determined by the distance $\rho(x, y)$ given a travel speed (the latter is fixed); and

$$t \longmapsto w_t : [0, T] \to X$$

is the specific (rectilinear, in our case) trajectory of the motion. The "regularity" mentioned in discussion of the use of (4.1) has the following sense: the point u is assumed to not to belong to the interval $[x; y] \triangleq \{\alpha x + (1 - \alpha) y : \alpha \in [0, 1]\}$. In absence of this regularity, the cost of travel from x to y is defined as a sufficiently large penalty constant. In definitions of the interior jobs, we follow the convention [11, Section 6] in determining the dose incurred by the agent during the local motion from the entry point (into the near zone of the radiation source; a megalopolis is its discretization) to the source itself; it is assumed that during the subsequent return travel to the exit point (from the near zone) there is no radiation from this source since it has been dismantled. This scheme is described in detail in [19, Section 4].

We use construction of [20, Section 6] for $\mathbf{c}, c_1, \ldots, c_N$. In connection with the construction of $\mathbf{c}$ we use [20, (6.17), (6.20), (6.39)]. For construction of $c_1, \ldots, c_N$, we use [20, (6.39)]. Of course, in [20, (6.17), 6.20), (6.39)], impact of the single source is considered. The function f is supposed identically equal zero. We recall also constructions of [21].

5. Software implementation and computational experiment

In this section, we describe the practical implementation of the procedure that constructs the Bellman function layers through independent computations by computational nodes and optimizes the starting point. Let us start by considering the implementation of the independent computations scheme. Assume that each layer connected with $K \in \mathfrak{C}_{N-1}$ is processed by several computational cores that share RAM. These cores together are called a computational node of the cluster; the latter is thus a union of computational nodes.

Data storage. Let us consider data storage on a single node of the cluster. For every set of objective points K, we may have to store the shortest paths set (SPS) (that is, the Bellman function layers) that pass through this set. Every shortest path in SPS differs from other paths in this SPS by its starting point and is the shortest among all other paths with the same starting point. An SPS may theoretically have as many paths as the cardinality of the corresponding set K, however, normally, there are less since not every point of this set can be initial in view of precedence constraints. In accordance with this, we store SPS in a hash table, with the aim of decreasing memory usage compared with an array. The key of the hash table is the bit mask of the set, where, if a bit at position i is set, then the point i is present in this set.

Main algorithm. The main node reads the input data from a file, which describes the objective sets that must be visited, the set of starting points, and the address pairs (precedence constraints). Next, the main node constructs the family

$$\mathfrak{C}_{N-1} = \{\overline{1,N} \setminus \{t\} : t \in \mathbf{I}(\overline{1,N})\},$$

every element of which is a cardinality $N-1$ set. The sets $K \in \mathfrak{C}_{N-1}$ are distributed between the nodes through the MPI protocol. Every node has k computational cores with shared RAM. At the node connected with the set $K \in \mathfrak{C}_{N-1}$, the Bellman function layers are shared between the cores in a uniform way. There is no exchange of data between the cores because the RAM is shared by all of them; the fragments of state space and the Bellman function layers are distributed between the cores of a single node through the OpenMP library. Then, in a parallel mode that conforms to the theoretical scheme [17, 18], which was implemented in [11] (see also [9, 10] for one-element megalopolises), the layers $v_1, \ldots, v_{N-1}$ of the "whole" (suitable for all the initial states from X^0) of the Bellman function are computed. After that, a relatively simple optimization procedure for

$$v_N(x, \overline{1,N}),$$

$x \in X^0$, is conducted, in the spirit of (3.5). This yields the grobal extremum $\mathbb{V}$ and the point $x^0 \in X^0$ with the property

$$V[x^0] = \mathbb{V}. \tag{5.1}$$

When an optimal solution in the form of a route-track pair is required, in addition to v_{N-1}, it is necessary to store all the Bellman function layers, which were determined by the main algorithm (when only the global extremum $\mathbb{V}$ and optimal initial state x^0 with property (5.1) are required, it is not necessary to store the mentioned Bellman function layers; it will suffice to have a procedure for constructing only the single layer v_{N-1} permitting the intermediary layers to be overwritten). Thus, assume $v_0, v_1, \ldots, v_{N-1}$ are known. Then, the main node constructs the optimal route for the flow where $V[x^0] = \mathbb{V}$ by means of finding the local extrema in a way similar to [7, 16, 19].

Computational experiment. In this section, we describe the solution of the routing problem on plane on the Uran supercomputer. The travel cost function is assumed to depend on the set of pending tasks; it is determined through relations similar to (4.1) (see also [20]); the function f is assumed to be zero since after all the megalopolises are visited and the corresponded interior jobs consisting of dismantling the radiating elements are conducted, the cost of return to base will be zero (since there is nothing to radiate anymore). Let us consider the case where the number of megalopolises is 48, i. e., $N = 48$ (in [11] a solution is given for the case of a significantly smaller dimension: it was assumed there $N = 30$ and $N = 31$). The megalopolises are contained inside circles on the plane. Thus, let the megalopolises, which imitate the entry and exit points to the spaces with radiation sources, be obtained by discretizing the circles (the boundaries of the near zones): on every circle, there are 30 equally spaced (in view of the angular distance). To every megalopolis, we assign a point object that imitates the radiation source in the space the megalopolis describes. The set of admissible starting points X^0 consists of 10 elements. In our example, the set $\mathbf{K}$ contains 45 address pairs that define the precedence constraints. Let this set have the following address pairs:
(38,45) (42,24) (22,7) (23,26) (32,17) (46,31) (34,8) (4,24) (17,8) (3,45) (0,26) (31,7) (3,20) (2,28) (18,47) (5,40) (36,25) (20,9) (7,6) (47,32) (46,40) (28,8) (33,5) (26,5) (0,34) (43,35) (9,27) (1,2) (1,37) (0,31) (7,23) (23,28) (39,31) (24,29) (17,45) (44,6) (29,11) (32,25) (2,14) (2,20) (15,36) (37,46) (21,10) (35,45) (12,37),

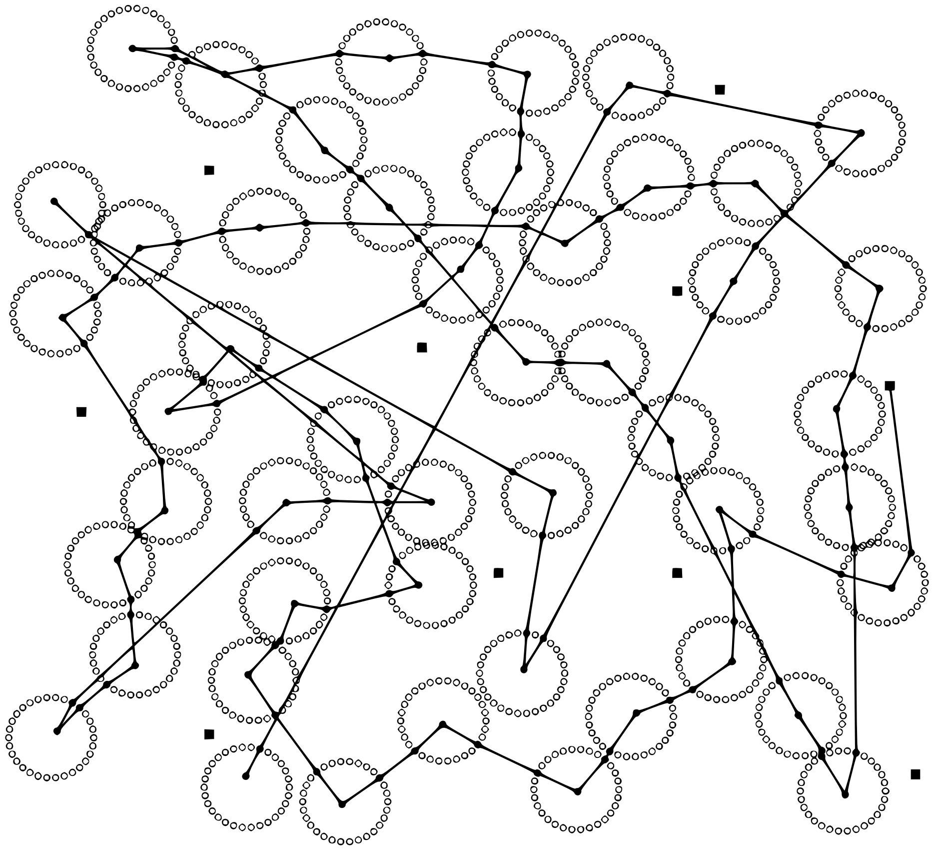


Figure 1. Route and track for visiting 48 megalopolises.

where the first argument specifies the sender and the second argument specifies the receiver. To verify the theoretical construction in practice, we implemented it in C++ for the Uran supercomputer. The program works under a 64-bit Linux operating system. The computational experiment was conducted on the nodes of the Uran cluster with the following characteristics:

two six-core Intel Xeon X5675 (3.07GHz) processors

192 GB RAM

2×12 MB Level 2 cache

8 Tesla M2090 GPUs (6 GB Global Memory)

400 GB local hard disk drive

The experiment used 20 cluster nodes, each of which had 12 cores. Thus, our practical implementation used 240 computational cores. The computations resulted in the starting point, route and track, see Fig. 1. The following results were obtained: $\mathbb{V} = 1.417074$ (extremum of the problem); the computation time was 15.772 seconds; the maximum RAM usage for a single computational node was 26.246 MB.

In a separate computational experiment, we considered the same problem with only 20 points per megalopolis (recall that those are viewed as exit/entry points into the facility associated with the megalopolis), equally spaced (with respect to angular distance). The following results were obtained: $\mathbb{V} = 1.4208160$ (extremum of the problem); the computation time was 13.256 seconds; the maximum RAM usage for a single computational node was 24.523 MB.

One could note that as the number of cities per megalopolis decreases, the extremum of the problem increases somewhat; this may be connected with the fact that in the second case there are fewer possible tracks, which, in its turn, makes the result worse.

In Fig. 1, the squares denote the admissible starting points. Transparent circles denote the cities in the megalopolises. Filled circles denote the entry and exit points of the megalopolises and the radiation sources inside them.

6. Computation with application of greedy algorithm

In this section we consider the solution of our basic problem by greedy algorithm similar to [33, Section 6] (in connection with construction of optimal algorithm on the base of DP, we note [7, 16, 19]). Now, we note only brief scheme of the clear greedy algorithm.

Namely, we fix $x^0 \in X^0$, suppose $\mathbf{z}^{(0)} \triangleq (x^0, x^0)$, and consider the problem

$$\mathbf{c}(x^0, \mathrm{pr}_1(z), \overline{1,N}) + c_j(z, \overline{1,N}) \longrightarrow \min, j \in \mathbf{I}(\overline{1,N}), z \in \mathbb{M}_j. \tag{6.1}$$

Now, we choose $\mathbf{j}_1 \in \mathbf{I}(\overline{1,N})$ and $\mathbf{z}^{(1)} \in \mathbb{M}_{\mathbf{j}_1}$ for which

$$\begin{gathered}\mathbf{c}(x^0, \mathrm{pr}_1(\mathbf{z}^{(1)}), \overline{1,N}) + c_{\mathbf{j}_1}(\mathbf{z}^{(1)}, \overline{1,N}) = \\ = \min_{j \in \mathbf{I}(\overline{1,N})} \min_{z \in \mathbb{M}_j} \big[\mathbf{c}\big(x^0, \mathrm{pr}_1(z), \overline{1,N}\big) + c_j(z, \overline{1,N})\big].\end{gathered} \tag{6.2}$$

Then, we obtain that

$$(\mathrm{pr}_2(\mathbf{z}^{(1)}), \overline{1,N} \setminus \{\mathbf{j}_1\}) \in D_{N-1}$$

Now, we have the above-mentioned position. Consider the problem

$$\mathbf{c}(\mathrm{pr}_2(\mathbf{z}^{(1)}), \mathrm{pr}_1(z), \overline{1,N} \setminus \{\mathbf{j}_1\}) + c_j(z, \overline{1,N} \setminus \{\mathbf{j}_1\}) \longrightarrow \min, j \in \mathbf{I}(\overline{1,N} \setminus \{\mathbf{j}_1\}), z \in \mathbb{M}_j.$$

We choose $\mathbf{j}_2 \in \mathbf{I}(\overline{1,N} \setminus \{\mathbf{j}_1\})$ and $\mathbf{z}^{(2)} \in \mathbb{M}_{\mathbf{j}_2}$ for which

$$\begin{gathered}\mathbf{c}(\mathrm{pr}_2(\mathbf{z}^{(1)}), \mathrm{pr}_1(\mathbf{z}^{(2)}), \overline{1,N} \setminus \{\mathbf{j}_1\}) + c_{\mathbf{j}_2}(\mathbf{z}^{(2)}, \overline{1,N} \setminus \{\mathbf{j}_1\}) = \\ = \min_{j \in \mathbf{I}(\overline{1,N} \setminus \{\mathbf{j}_1\})} \min_{z \in \mathbb{M}_j} \big[\mathbf{c}\big(\mathrm{pr}_2(\mathbf{z}^{(1)}), \mathrm{pr}_1(z), \overline{1,N} \setminus \{\mathbf{j}_1\}\big) + c_j(z, \overline{1,N} \setminus \{\mathbf{j}_1\})\big].\end{gathered} \tag{6.3}$$

Then, we obtain the next inclusion

$$(\mathrm{pr}_2(\mathbf{z}^{(2)}), \overline{1,N} \setminus \{\mathbf{j}_1; \mathbf{j}_2\}) \in D_{N-2}$$

The further construction are realized similar to (6.2) and (6.3) up to exhaustion of all list $\overline{1,N}$. We obtain two next finite processions

$$(\mathbf{j}_k)_{k \in \overline{1,N}} : \overline{1,N} \longrightarrow \overline{1,N},$$

$$(\mathbf{z}^{(k)})_{k \in \overline{0,N}} : \overline{0,N} \longrightarrow \mathbb{X} \times \mathbb{X}.$$

In addition, $\mathbf{i}[x^0] \triangleq (\mathbf{j}_k)_{k \in \overline{1,N}} \in \mathbf{A}$ and $(\mathbf{z}^{(k)})_{k \in \overline{0,N}} \in \mathcal{Z}_{\mathbf{i}[x^0]}[x^0]$. Of course, the value

$$\mathfrak{C}_{\mathbf{i}[x^0]}[(\mathbf{z}^{(k)})_{k \in \overline{0,N}}] \in \mathbb{R}_+ \tag{6.4}$$

corresponds to our initial state $x^0 \in X^0$. Therefore, we introduce designation

$$w[x^0] \triangleq \mathfrak{C}_{\mathbf{i}[x^0]}[(\mathbf{z}^{(k)})_{k \in \overline{0,N}}].$$

The analogous constructions are realized for all $x \in X^0$. As a result, we obtain values

$$w[x], x \in X^0.$$

We choose $x_0 \in X^0$ by the rule

$$w[x_0] = \min_{x \in X^0} w[x] \tag{6.5}$$

We consider (6.5) as upper estimate for $\mathbb{V}$ and use $\mathbf{i}[x_0]$ as the solution corresponding to greedy algorithm.

Consider a variant of computation. We preserve parameters of Section 6: $N = 48$, $|M_j| = 30$, $|\mathbf{K}| = 45$ (concrete address pairs are indicated in Section 5), $|X^0| = 10$. Under computation with employment of greedy algorithm, the result value

$$\min_{x \in X^0} w[x] = 1.884528 \tag{6.6}$$

was obtained. The minimizing point x^0 (see (6.5)) coincides with (103.12; 5.06). In this connection, we recall that, for optimal solution (see Section 5), we have $x^0 = (100.00; 53.26)$, for best initial state. In addition, for extremes realized by optimal and greedy algorithms, we obtain the following ratio: global extremum achievable by the DP procedure improves the value (6.6) about 25%. Of course, time of computing under employment of greedy algorithm about 173 seq. (recall that analogous time for optimal algorithm is 15772 seq.). We note that our greedy algorithm can be used for solving of problems having big detention. This algorithm was used in problem with 254 megalopolises and $|\mathbf{K}| = 45$. In this case, the value (6.5) and point x^0 were obtained during 1687 seq (most of this time was spent on calculating the cost function).

7. Conclusion

In this paper, we consider the issues related to the solving a routing problem with precedence constraints and complicated travel cost functions aimed at applications connected with conducting a sequence of actions in a high-radiation area. However, similar problem statements are also present in other applications. For example, in particular, a “more complex” general statement can be used to solve a problem connected with CNC plate cutting machines; see, in particular, [22–28]. A comparison with the latter is natural: both statements are very much oriented towards the practice and conduct routing with “interior” tasks. Travel cost functions’ dependence on the set of pending tasks can be connected with the need to account for various constraints of dynamic character (see, [16]), specifically, a system of penalties. In this problem, the starting point is normally known in advance — if we consider the engineering problems connected with nesting; however, thinking in perspective, it may be worthwhile to consider statement (2.10) as a way of tackling the problem of choosing the initial state of the tool. This may be of importance in view of the characteristic constraints (rigidity of the whole plate and each item). In connection with TSP and TSP-like problems, let us note [29] and [30] concerned with two versions of dynamic programming and [31], which deals with the branch-and-bound method. In connection with construction of production-oriented heuristics, note [32]. However, it appears that real-life problems connected with routing have many specific issues and peculiarities, and must thus be treated with special methods (first and foremost, special heuristics); in this paper, we have endeavored to construct some.

REFERENCES

1. Melamed I. I., Sergeev S. I., Sigal I. The traveling salesman problem. Issues in theory. *Autom. Remote Control*, 1989. Vol. 50, No. 9. P. 1147–1173.
2. Melamed I. I., Sergeev S. I., Sigal I. The traveling salesman problem. Exact methods. *Autom. Remote Control*, 1989. Vol. 50, No. 10. P. 1303–1324.
3. Melamed I. I., Sergeev S. I., Sigal I. The traveling salesman problem. Approximate algorithms. *Autom. Remote Control*, 1989. Vol. 50, No. 11. P. 1459–1479.

4. Gutin G., Punnen A. P. *The Traveling Salesman Problem and Its Variations.* New York: Springer, 2002. DOI: 10.1007/b101971
5. Cook W. J. *In Pursuit of the Traveling Salesman. Mathematics at the Limits of Computation.* New Jersey: Princeton University Press, 2012. p. 248.
6. Gimadi E. Kh., Khachai M. Yu. *Ekstremalnye zadachi na mnozhestvax perestanovok* [Extremal Problems on Sets of Permutations], Yekaterinburg: UMC UPI, 2016. p. 220 (in Russian)
7. Chentsov A. G., Chentsov A. A. Route problem with constraints depending on a list of tasks. *Doklady Mathematics*, 2015. Vol. 92, No. 3. P. 685–688. DOI: 10.1134/S1064562415060083
8. Chentsov A. G., Chentsov A. A. A discrete-continuous routing problem with precedence conditions. *Proc. Steklov Inst. Math.*, 2018. Vol. 300, No. 1. P. 56–71. DOI: 10.1134/S0081543818020074
9. Chentsov A. G., Grigoryev A. M. Dynamic Programming Method in a Routing Problem: a Scheme of Independent Computations. *Mekhatronika, Avtomatizatsiya, Upravlenie*, 2016. Vol. 17, No. 12. P. 834–846. DOI: 10.17587/mau.17.834-846 (in Russian)
10. Chentsov A. G., Grigoryev A. M. A scheme of independent calculations in a precedence constrained routing problem. *Lecture Notes in Computer Science*, Vol. 9869: Intern. Conf. on Discrete Optimization and Operations Research (DOOR–2016), 2016. P. 121–135. DOI: 10.1007/978-3-319-44914-2_10
11. Chentsov A. G., Grigoryev A. M., Chentsov A. A. Decommissioning of nuclear facilities: minimum accumulated radiation dose routing problem. *CEUR-WS Proc.*, Vol. 1987: 8th Intern. Conf. on Optimization and Applications (OPTIMA–2017), 2017. P. 123–130. http://ceur-ws.org/Vol-1987/paper19.pdf
12. Chentsov A. G., Khachai M. Y., Khachai D. M. An exact algorithm with linear complexity for a problem of visiting megalopolises. *Proc. Steklov Inst. Math.*, 2016. Vol. 295, supp. 1. P. 38–46. DOI: 10.1134/S0081543816090054
13. Korobkin V. V., Sesekin A. N., Tashlykov O. L., and Chentsov A. G. *Metody marshrutizacii i ih prilozheniya v zadachah povysheniya ehffektivnosti i bezopasnosti ehkspluatacii atomnyh stancij* [Methods of Routing with Application to the Problems of Safety Enhancement and Operational Effectiveness of Nuclear Power Plants], Ed. I.A. Kalyaev. Moscow: Novye Tekhnologii, 2012. (in Russian)
14. Kuratowski K., Mostowski A. *Set Theory.* Amsterdam: North-Holland Publishing Company, 1968. p. 417.
15. Dieudonné J. *Foundations of Modern Analysis.* New York: Academic Press, 1969. 407 p.
16. Chentsov A. G., Chentsov P. A. Routing under constraints: Problem of visit to megalopolises. *Autom. Remote Control*, 2016. Vol. 77, No. 11. P. 1957–1974. DOI: 10.1134/S0005117916110060
17. Chentsov A. G. On a parallel procedure for constructing the Bellman function in the generalized problem of courier with internal jobs. *Autom. Remote Control*, 2012. Vol. 73, No. 3. P. 532–546. DOI: 10.1134/S0005117912030113
18. Chentsov A. G. A parallel procedure of constructing Bellman function in the generalized courier problem with interior works. *Vestnik YuUrGU. Ser. Mat. Model. Progr.*, 2012. No. 18. P. 53–76. (in Russian)
19. Chentsov A. A., Chentsov A. G., and Chentsov P. A. Elements of Dynamic Programming in the Extremal Problems of Routing. *Autom. Remote Control*, 2014. Vol. 75, No. 3. P. 537–550 DOI: 10.1134/S0005117914030102
20. Chentsov A. G., Chentsov A. A. A model variant of the problem about radiation sources utilization (iterations based on optimization insertions). *Izv. Inst. Mat. Inform. Udmurt. Gos. Univ.*, 2017. Vol. 50. P. 83–109. DOI: 10.20537/2226-3594-2017-50-08 (in Russian)
21. Chentsov A. G., Chentsov A. A., Grigoryev A. M. On one routing problem modeling movement in radiation fields. *Vestn. Udmurtsk. Univ. Mat. Mekh. Komp. Nauki*, 2017. Vol. 27, No. 4. P. 540–557. DOI: 10.20537/vm170405 (in Russian)
22. Petunin A.A. About some strategy of formation of a route of the cutting tool by development of the controlling programs for the thermal sheet cutting machines. *The UGATU Bulletin. Series: Control, ADP equipment and informatics*, 2009. Vol. 13, No. 2 (35). P. 280–286.
23. Frolovskii V.D. Computer-aided design of the control programs for thermal metal cutting on NPC machines. *Informacionnye tekhnologii v proektirovanii i proizvodstve*, 2005. No. 4. P. 63–66.
24. Wang G.G. and Xie S.Q. Optimal process planning for a combined punch-and-laser cutting machine using ant colony optimization. *Int. J. Product. Res.*, 2005. Vol. 43, No. 11. P. 2195–2216. DOI: 10.1080/00207540500070376

25. Lee M.-K. and Kwon K.-B. Cutting path optimization in NC cutting processes using a two-step genetic algorithm. *Int. J. Product. Res.*, 2006. Vol. 44, No. 24. P. 5307–5326. DOI: 10.1080/00207540600579615
26. Jing Y. and Zhige C. An optimized algorithm of numerical cutting-path control in garment manufacturing. *Adv. Mater. Res.*, 2013. Vol. 796. P. 454–457. DOI: 10.4028/www.scientific.net/AMR.796.454
27. Ganelina N.D. and Frolovskii V.D. On constructing the shortest circuits on a set ofline segments. *Sib. Zh. Vychisl. Mat.* [Siberian J. of Numer. Mathematics], 2006. Vol. 9, No. 3. P. 241–252.
28. Verkhoturov M.A. and Tarasenko P.Yu. Software for the problems of optmization of the cutting tool path for planar figure cutting on the basis of chain cutting. *Vestn. UGATU*, 2008. Vol. 10, No. 2 (27). P. 123–130. http://journal.ugatu.ac.ru/index.php/Vestnik/article/view/1274/1103
29. Bellman R. Dynamic programming treatment of the travelling salesman problem. *J. ACM.* 1962. Vol. 9, No. 1. P. 61–63. DOI: 10.1145/321105.321111
30. Held M. and Karp R.M. A dynamic programming approach to sequencing problems. *J. Soc. Ind. Appl. Math.*, 1962. No. 10 (1). P. 196–210. DOI: 10.1137/0110015
31. Little J. D. C., Murti K. G., Sweeney D. W., and Karel C. Algorithm for the traveling salesman problem. *Econ. Mat. Metod.*, 1965. Vol. 1, No. 1. P. 94–107.
32. Escudero L. An inexact algorithm for the sequential ordering problem. *Eur. J. Oper. Res.*, 1988. Vol. 37, No. 2. P. 236–249.
33. Chentsov A.G., Chentsov A.A., Chentsov P.A. Extremal routing problem with internal losses. *Proc. Steklov Inst. Math.* 2009. Vol. 264, suppl. 1. P. 87–106. DOI: 10.1134/S0081543809050071

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 56–68

DOI: 10.15826/umj.2018.2.007

A STABLE METHOD FOR LINEAR EQUATION IN BANACH SPACES WITH SMOOTH NORMS[1]

Andrey A. Dryazhenkov† and Mikhail M. Potapov††

Lomonosov Moscow State University,
Leninskie Gory, Moscow, Russia, 119991

†andrja@yandex.ru, ††mmpotapovrus@gmail.com

Abstract: A stable method for numerical solution of a linear operator equation in reflexive Banach spaces is proposed. The operator and the right-hand side of the equation are assumed to be known approximately. The corresponding error levels may remain unknown. Approximate operators and their conjugate ones must possess the property of strong pointwise convergence. The exact normal solution is assumed to be sourcewise representable and some upper estimate for the norm of its source element must be known. The norm in the Banach space of solutions is supposed to satisfy the following smoothness-type condition: some function of the norm must be differentiable. Under these conditions a stability of the method with respect to nonuniform perturbations in operator is shown and the strong convergence to the normal solution is proved. A boundary control problem for the one-dimensional wave equation is considered as an example of possible application. The results of the model numerical experiments are presented.

Keywords: Linear operator equation, Banach space, Numerical solution, Stable method, Sourcewise representability, Wave equation.

Introduction

The problem of finding solution to a linear operator equation arises in many fields of applied mathematics when solving integral equations, some boundary value problems, systems of linear equations and other linear inverse problems. The known complication that can arise thereby is the ill-posedness of such inverse problems. This means that small changes in initial data (coefficients of the system of linear equations, the right-hand sides of equations, boundary data, coefficients of differential operator, etc.) can cause loss of existence or uniqueness of the perturbed problem solution or lead to not small changes in this solution. To deal with the issues of such types many regularization methods were proposed: Tikhonov regularization method [23, 24], residual method [17], method of quasi-solutions [12], residual principle [16], iterative regularization methods [2] and many others [3, 10, 21, 22, 25]. Most of them require knowledge of error levels in initial data approximation or knowledge of some compact set containing a sought solution. In many applications these assumptions are rather hard to be ensured. Instead of these traditional assumptions our method requires a sought solution to be sourcewise representable and, moreover, some majorant for the source norm to be known. It allows anyone who wants to apply the method to focus on researching corresponding properties of the exact problem.

In this paper we consider a linear operator equation

$$\mathcal{A}u = f \tag{1}$$

in reflexive Banach spaces H and F, where $\mathcal{A} \in \mathcal{L}(H \to F)$ is a linear bounded operator and $f \in F$ is a given element. It is required to find normal solution u_*, i. e. a solution u_* to (1) with a minimal

[1]This work was supported by the grant of Russian Science Foundation (project 14-11-00539).

norm in the space H:

$$u_* = \arg\min_{u \in U} \|u\|_H, \quad U = \{u \in H \,|\, \mathcal{A}u = f\}. \tag{2}$$

In the sequel the norm of the space H will be supposed to be strictly convex, so the solution u_* to the problem (2) is unique, and it exists if equation (1) has a solution [8, Proposition 1.2, p. 35].

Suppose that instead of exact data $\mathcal{A}$ and f some of their approximations $\mathcal{A}_n \in \mathcal{L}(H \to F)$ and $f_n \in F$, $n = 1, 2 \ldots$, are known. The asymptotic properties of the method will be studied under the condition that the approximate data converge to the exact ones in the following sense:

$$\begin{gathered} \|\mathcal{A}_n u - \mathcal{A}u\|_F \to 0, \quad \forall\, u \in H, \quad \|\mathcal{A}_n^* v - \mathcal{A}^* v\|_{H^*} \to 0, \quad \forall\, v \in F^*, \\ \|f_n - f\|_F \to 0 \quad \text{as} \quad n \to \infty. \end{gathered} \tag{3}$$

Here and below $\mathcal{A}^* : F^* \to H^*$ and $\mathcal{A}_n^* : F^* \to H^*$ are operators adjoint to $\mathcal{A}$ and $\mathcal{A}_n$. Note that the first two limit relations in (3) are weaker than conditions of uniform convergence usually required in the traditional regularizing procedures [3, 10, 16], [22]–[25]. Also we do not require in (3) the knowledge of any error levels.

A stable method of solving the problem (2) under perturbations of type (3) in Hilbert spaces H and F was proposed in [19]. Briefly recall this method for the convenience of comparison. In [19] the following basic assumptions were accepted:

H1. Spaces H and F are Hilbert and identified with their adjoint spaces in the Riesz sense: $H \simeq H^*$, $F \simeq F^*$.

H2. Equation (1) has a solution.

H3. The solution u_* to (2) is sourcewise representable: $u_* \in R(\mathcal{A}^*)$, where $R(\mathcal{A}^*)$ denotes range of operator $\mathcal{A}^* : F \to H$. It means that there exists a source element $v_* \in F$ such that $u_* = \mathcal{A}^* v_*$.

H4. Some majorant r_* of the source norm is known: $\|v_*\|_F \le r_*$.

It is well-known that the solution u_* to (2) belongs to the closure of $R(\mathcal{A}^*)$ [10, Proposition 2.3, p. 33], so the assumption H3 is rather natural and holds true for any operator $\mathcal{A}$ with closed range.

The method from [19] is then formulated as follows: find a solution $v_n \in F$ to the following quadratic optimization problem

$$\begin{gathered} I_n(v_n) \le \inf_{v \in V} I_n(v) + \varepsilon_n, \quad \varepsilon_n \ge 0, \\ V = \{v \in F \,|\, \|v\|_F \le r_*\}, \quad I_n(v) = \frac{1}{2}\, \|\mathcal{A}_n^* v\|_H^2 - \langle v, f_n \rangle_F, \end{gathered} \tag{4}$$

and set element $u_n = \mathcal{A}_n^* v_n$ as a final approximation for the sought solution to (2). Here $\langle \cdot, \cdot \rangle_F$ denotes the inner product in space F.

Theorem 1 [19]. *Let assumptions* (3), H1–H4 *be fulfilled, let u_n be an output of the described method and $\varepsilon_n \to 0$ as $n \to \infty$. Then the convergence $\|u_n - u_*\|_H \to 0$ holds true.*

The method proposed below is an extension of the described method from [19] to Banach spaces with smoothness-type property of the norm in space H.

The rest of the paper is organized as follows. In the next Section 2, we formulate some assumptions about the spaces H, F and the special properties of the exact solution. In Section 3 the method is described, and in the Section 4 its stability is proved. In Section 5 one of the possible applications to the boundary control problem for the 1-D wave equation is considered, and in final Section 6 corresponding numerical results are provided.

1. Basic Assumptions and Auxiliary Statements

The method presented in the next section for Banach spaces requires the following assumptions:

B1. H and F are reflexive Banach spaces.

B2. The norms in H and H^* are strictly convex.

B3. For the norm in H^* the Radon–Riesz property holds true: if sequence $\{g_n\} \subset H^*$ converges weakly to $g_0 \in H^*$: $g_n \xrightarrow{w} g_0$, and the corresponding sequence of norms also converges: $\|g_n\|_{H^*} \to \|g_0\|_{H^*}$, then the sequence $\{g_n\}$ converges strongly: $\|g_n - g_0\|_{H^*} \to 0$.

B4. Let $\phi \in C[0, +\infty[$ be a continuous strictly increasing function, $\phi(0) = 0$, $\phi(+\infty) = +\infty$ and let ϕ^{-1} be inverse of ϕ. Let two functionals $P : H \to \mathbb{R}$ and $K : H^* \to \mathbb{R}$ are defined as

$$\begin{aligned} P(u) = p(\|u\|_H), \quad p(x) = \int_0^x \phi(\xi)\, d\xi, \\ K(g) = k(\|g\|_{H^*}), \quad k(x) = \int_0^x \phi^{-1}(\xi)\, d\xi. \end{aligned} \tag{1.1}$$

These functional are assumed to be Fréchet differentiable: $P \in C^1(H)$, $K \in C^1(H^*)$.

B5. Equation (1) has a solution.

B6. The solution u_* to (2) is sourcewise representable in the following sense: there exists an element $v_* \in F^*$ such that $u_* = \mathcal{J}_H \mathcal{A}^* v_*$, where mapping $\mathcal{J}_H : H^* \to H$ is defined as

$$\mathcal{J}_H g = K'(g), \quad \forall g \in H^*. \tag{1.2}$$

B7. Some majorant r_* of the source norm is known: $\|v_*\|_{F^*} \le r_*$.

Remark 1. Using reflexivity of H and Asplund's duality mapping representation theorem [5, Theorem 4.4, p. 26], it is not hard to see that $\mathcal{J}_H$ defined in (1.2) is in fact duality mapping with weight (or gauge) function $\phi^{-1}(x)$.

Let us explain the meaning of the assumption B6. As in the case of Hilbert spaces H and F, this assumption is fulfilled for operators $\mathcal{A}$ with closed range. The corresponding proof will be presented now.

Theorem 2. *Let assumptions* B1, B2, B4, B5 *be fulfilled and* $\mathcal{A}u_* = f$*. Then* u_* *is solution to* (2) *if and only if*

$$P'(u_*) \in \overline{R(\mathcal{A}^*)}, \tag{1.3}$$

where $\overline{R(\mathcal{A}^*)}$ *is closure of* $R(\mathcal{A}^*)$*.*

P r o o f. Let u_* be a solution to (2). Let us prove that (1.3) takes place. Consider the following minimization problem:

$$P(u) \to \min, \quad \mathcal{A}u = f. \tag{1.4}$$

Since function $p(x)$ is strictly increasing and $P(u) = p(\|u\|_H)$, this problem is equivalent to (2), and the element u_* is the unique solution to (1.4). Also consider linear auxiliary minimization problem:

$$\langle P'(u_*), u\rangle \to \inf, \quad \mathcal{A}u = 0. \tag{1.5}$$

Here and below, the expression $\langle f, u\rangle$ is understood as the value of linear continuous functional $f \in H^*$ on the element $u \in H$. Notice that optimal value of minimizing functional in (1.5) is nonnegative. Indeed, if there exists an element $\widehat{u} \in H$ such that $\langle P'(u_*), \widehat{u}\rangle < 0$ and $\mathcal{A}\widehat{u} = 0$, then we can consider elements $u_\alpha = u_* + \alpha\widehat{u}$, $\alpha \geq 0$. Using definition of Fréchet derivative we get

$$P(u_\alpha) = P(u_*) + \alpha\langle P'(u_*), \widehat{u}\rangle + o(\alpha),$$

where $o(\alpha)/\alpha \to 0$ as $\alpha \to 0$. It means that for all sufficiently small $\alpha > 0$ $P(u_\alpha) < P(u_*)$ and $\mathcal{A}u_\alpha = \mathcal{A}u_* + \alpha\mathcal{A}\widehat{u} = \mathcal{A}u_* = f$, so u_* is not the solution to (1.4). This contradiction shows that $\langle P'(u_*), u\rangle \geq 0$ for all $u \in H$ such that $\mathcal{A}u = 0$, i.e. for all $u \in N(\mathcal{A})$, where $N(\mathcal{A})$ denotes the kernel of $\mathcal{A}$. Since the kernel $N(\mathcal{A})$ is a linear subspace of H, it means that $\langle P'(u_*), u\rangle = 0$ for all $u \in N(\mathcal{A})$. In other words, we have

$$P'(u_*) \in (N(\mathcal{A}))^\perp, \quad (N(\mathcal{A}))^\perp = \{g \in H^* \mid \langle g, u\rangle = 0,\ \forall u \in N(\mathcal{A})\}, \tag{1.6}$$

and equality $(N(\mathcal{A}))^\perp = \overline{R(\mathcal{A}^*)}$ (see [13, Theorem 1*, p. 357], using reflexivity of H) allows us to pass from (1.6) to (1.3).

On the other hand, let u_* be a solution to (1) and let inclusion (1.3) be fulfilled. We want to prove that $u_* = \widehat{u}$, where $\widehat{u}$ is a solution of (2). Let us suppose that $u_* \neq \widehat{u}$. Notice that under assumptions B2, B4 operator $P'(u)$ is strictly monotonic and that is why the following inequality holds true:

$$\langle P'(u_*) - P'(\widehat{u}), u_* - \widehat{u}\rangle > 0. \tag{1.7}$$

It was proved above that $\widehat{u}$ satisfies the condition (1.3), therefore $P'(u_*) - P'(\widehat{u}) \in \overline{R(\mathcal{A}^*)}$. With inequality (1.7) it implies that there exists an element $\widehat{v} \in F^*$ such that $\langle \mathcal{A}^*\widehat{v}, u_* - \widehat{u}\rangle > 0$, but $\langle \mathcal{A}^*\widehat{v}, u_* - \widehat{u}\rangle = \langle \widehat{v}, \mathcal{A}u_* - \mathcal{A}\widehat{u}\rangle = \langle \widehat{v}, f - f\rangle = 0$. This contradiction means that our assumption $u_* \neq \widehat{u}$ is not true, so u_* is indeed a solution to (2). □

Lemma 1. *Let assumptions* B1, B2, B4 *be fulfilled. Then* $K'(P'(u)) = u$, $\forall u \in H$ *and* $P'(K'(g)) = g$, $\forall g \in H^*$.

P r o o f. Let us extend functions $p(x)$ and $k(x)$ defined in (1.1) to the region $x \leq 0$ in the even way: $k(x) = k(-x)$, $p(x) = p(-x)$. Then these extensions will be convex dual. Indeed, for all $x \in \mathbb{R}$ concave function $xy - k(y)$ of variable y attains its maximum when $x - k'(y) = 0$, i.e. $\phi^{-1}(y) = x$. It means that

$$\sup_{y\in\mathbb{R}} (xy - k(y)) = x\phi(x) - k(\phi(x)) = x\phi(x) - \int_0^{\phi(x)} \phi^{-1}(\xi)\, d\xi. \tag{1.8}$$

Note that the following equality takes place for any strictly increasing smooth function $\psi \in C^1(\mathbb{R})$:

$$x\psi(x) - \int_0^{\psi(x)} \psi^{-1}(\xi)\, d\xi = x\psi(x) - \int_0^x \chi\psi'(\chi)\, d\chi = \int_0^x \psi(\chi)\, d\chi. \tag{1.9}$$

Passing in (1.9) to the limit as $\psi \to \phi$, $\psi^{-1} \to \phi^{-1}$ uniformly on any segment $[a, b]$, we obtain from (1.8) that

$$k^*(x) = \sup_{y\in\mathbb{R}} (xy - k(y)) = \int_0^x \phi(\chi)\, d\chi = p(x), \quad \forall x \in \mathbb{R},$$

so $k^*(x) = p(x)$. Applying Fenchel–Moreau theorem [8, Proposition 4.1, p.18] we get $k^{**} = p^* = k$, so functions k and p are dual. Then we get the duality of functions $P(u) = p(\|u\|_H)$ and

$K(g) = k(\|g\|_{H^*})$ (see [8, Proposition 4.2, p.19]). Finally, we pass to the lemma statement using the relation between subgradients of dual functions [8, Corollary 5.2, p. 22]. $\square$

Applying lemma 1 to (1.3) and using notation (1.2) we get the main result concerning assumption B6.

Corollary 1. *Let assumptions* B1, B2, B4, B5 *be fulfilled and let* u_* *be a solution to* (1): $\mathcal{A}u_* = f$. *Then* u_* *is solution to* (2) *if and only if*

$$u_* \in \mathcal{J}_H \overline{R(\mathcal{A}^*)}. \tag{1.10}$$

It means that assumption B6 is fulfilled for all operators $\mathcal{A}$ with closed range. For other operators this assumption contains additional requirement to the normal solution u_*, but it is rather close to the necessary condition (1.10).

2. Description of the Method

The algorithm proposed below in Banach spaces to find the normal solution (2) to the equation (1) in case of approximate data $\mathcal{A}_n$, f_n, $n = 1, 2, \ldots,$ is similar to its Hilbert version (4) from [19].

1. For the fixed sequence number n find an element $v_n \in V$ that satisfies the conditions

$$\begin{gathered} I_n(v_n) \leq \inf_{v \in V} I_n(v) + \varepsilon_n, \\ V = \{v \in F^* \mid \|v\|_{F^*} \leq r_*\}, \quad I_n(v) = K\big(\mathcal{A}_n^* v\big) - \langle v, f_n \rangle, \end{gathered} \tag{2.1}$$

   where r_* is taken from assumption B7 and $\varepsilon_n \geq 0$ is a parameter that allows to solve the optimization problem $I_n(v) \to \inf,\ v \in V$ approximately.

2. Set $u_n = \mathcal{J}_H \mathcal{A}_n^* v_n$ as an approximate solution to (2).

Remark 2. Note that for Hilbert spaces H and F we can take $\phi(x) = \phi^{-1}(x) = x$. Then $K(u) = P(u) = \|u\|_H^2/2$ and $\mathcal{J}_H u = K'(u) = u$. In this case method (2.1) fully coincides with the method from [19].

Remark 3. As in [19] instead of V we can use in (2.1) sets

$$V_n = \{v \in F_n^* \mid \|v\|_{F^*} \leq r_*\},$$

where F_n^* is a closed subspace of F^* such that $\mathcal{A}_n^* F_n^* = R(\mathcal{A}_n^*)$. In this case the proof of the method convergence does not change. For finite-dimensional approximate operators $\mathcal{A}_n$ and $\mathcal{A}_n^*$ which are usually used in practical computations, it makes possible to choose finite-dimensional subspaces F_n^* for variations of sources v. In this case, problem $I_n(v) \to \inf,\ v \in V_n$ turns into a finite-dimensional problem of minimization a smooth convex function $I_n(v)$ on a ball V_n. Note that ball is one of the simpliest convex closed bounded set with a non-empty interior. For an approximate solution of such problems, more precisely, an approximate solution by the value of the function, there is a well-developed arsenal of numerical methods.

3. Proof of Convergence

Let us examine the behavior of the approximate solutions u_n when perturbed data $\mathcal{A}_n$, f_n asymptotically approach their exact values $\mathcal{A}$, f in the sense of (3). To do this we need the following equivalent reformulation of the problem (2).

Lemma 2. *Let assumptions* B1, B2, B4–B6 *be fulfilled. Then an element* $u_* \in H$ *is the solution to* (2) *if and only if it can be represented as* $u_* = \mathcal{J}_H \mathcal{A}^* \widehat{v}$*, where* $\widehat{v}$ *is a solution to the following optimization problem:*

$$K\left(\mathcal{A}^* v\right) - \langle v, f\rangle \to \min, \quad v \in F^*. \tag{3.1}$$

P r o o f. The problem (3.1) is a smooth and convex one without constraints, so it is equivalent to finding an element $v \in F^*$ on which the derivative of the functional vanishes [8, Proposition 2.1, p. 36]:

$$\mathcal{A} K'\left(\mathcal{A}^* v\right) - f = 0.$$

Taking into account (1.2), this equation is equivalent to a system of two equations for the unknowns $(u, v) \in H \times F^*$:

$$\mathcal{A} u = f, \quad u = \mathcal{J}_H \mathcal{A}^* v. \tag{3.2}$$

Let u_* be a solution to (2). Then it follows from assumption B6 that u_* satisfies (3.2). On the other hand, if u_* satisfies (3.2) then using corollary 1 we get that u_* is the solution to (2). $\square$

Now we are ready to prove convergence of the method.

Theorem 3. *Let assumptions* B1–B7 *and conditions* (3) *be fulfilled. Let* u_n *be a final output of the method described above and* $\varepsilon_n \to 0$ *as* $n \to \infty$*. Then* $\|u_n - u_*\|_H \to 0$ *as* $n \to \infty$*.*

P r o o f. Space F^* is reflexive, and set V defined in (2.1) is convex, bounded and closed, therefore family $v_n \in V$ has in F^* a weak limit point $v_0 \in V$ [7, V. 4.7, p. 425]: $v_{n_m} \overset{w}{\to} v_0$ as $m \to \infty$. In order to simplify notation, we will omit symbol m from the subsequence n_m and write $v_n \overset{w}{\to} v_0$, $n \to \infty$. Then due to the strong pointwise convergence of $\mathcal{A}_n$ to $\mathcal{A}$ we have

$$\mathcal{A}_n^* v_n \overset{w}{\to} \mathcal{A}^* v_0. \tag{3.3}$$

Functional $K(g)$ is convex and continuous, hence it is weakly lower semicontinuous [9, Proposition 5, p. 74]. Denote $I(v) = K(\mathcal{A}^* v) - \langle v, f\rangle$ and notice that the following inequalities are valid:

$$I(v_0) \le \underline{\lim}\, I_n(v_n) \le \overline{\lim}\, I_n(v_n) \le \overline{\lim}\left(I_n(v_0) + \varepsilon_n\right) = I(v_0). \tag{3.4}$$

The first inequality in (3.4) is due to weak lower semicontinuity of $K(g)$ and strong convergence $\|f_n - f\|_F \to 0$. The third inequality follows from (2.1) and the inclusion $v_0 \in V$. The equality is due to (3). From (3.4) it follows that there exists $\lim I_n(v_n) = I(v_0)$, i. e.

$$K\left(\mathcal{A}_n^* v_n\right) - \langle v_n, f_n\rangle \to K(\mathcal{A}^* v_0) - \langle v_0, f\rangle.$$

Since $\langle v_n, f_n\rangle \to \langle v_0, f\rangle$ we also have

$$K\left(\mathcal{A}_n^* v_n\right) \to K(\mathcal{A}^* v_0), \quad \|\mathcal{A}_n^* v_n\|_{H^*} \to \|\mathcal{A}^* v_0\|_{H^*}, \tag{3.5}$$

where the last convergence takes place because the function $k(x)$ is strictly increasing. Using (3.3), (3.5) and Radon–Riesz property of norm from assumption B3, we get strong convergence

$$\|\mathcal{A}_n^* v_n - \mathcal{A}^* v_0\|_{H^*} \to 0. \tag{3.6}$$

Since source element v_* from assumption B6 belongs to V, it follows from (2.1) that $I_n(v_n) \le I_n(v_*) + \varepsilon_n$. Passing to a limit and using (3.6) and convergence $\langle v_n, f_n\rangle \to \langle v_0, f\rangle$, we get $I(v_0) \le I(v_*)$. Lemma 2 states that v_* is a solution to global optimization problem (3.1). That is why

$$I(v_0) = I(v_*) \le I(v), \quad \forall v \in F^*.$$

This means that v_0 is also a solution to the problem (3.1), so using lemma 2 once more, but in opposite direction, we obtain that the only solution u_* to the problem (2) can be represented by the source v_0:

$$u_* = \mathcal{J}_H \mathcal{A}^* v_0.$$

Assumption B4 implies strong continuity of $\mathcal{J}_H$, which with (3.6) leads to the limit relation

$$\|u_n - u_*\|_H = \|\mathcal{J}_H \mathcal{A}_n^* v_n - \mathcal{J}_H \mathcal{A}^* v_0\|_H \to 0. \tag{3.7}$$

Notice that our proof holds true for all weak limit points $v_0 \in V$, and that is why convergence (3.7) is valid for arbitrary family of approximate data $\mathcal{A}_n, f_n$ possessing asymptotic properties (3). □

4. Application to the Boundary Control Problem

In order to illustrate the application ability of the method, consider the following model boundary control problem for one-dimensional wave equation:

$$\begin{gathered} y_{tt}(t,x) = y_{xx}(t,x), \quad (t,x) \in (0,T) \times (0,l), \\ y|_{x=0} = u(t), \quad y|_{x=l} = 0, \quad t \in (0,T), \\ y|_{t=0} = 0, \quad y_t|_{t=0} = 0, \quad x \in (0,l). \end{gathered} \tag{4.1}$$

The goal of control actions $u(t)$ is to drive the system to a given final state $f(x) = (f^0(x), f^1(x))$ at a given time $T \ge 2l$:

$$y|_{t=T} = f^0(x), \quad y_t|_{t=T} = f^1(x), \quad x \in (0,l). \tag{4.2}$$

The spaces H and F of controls $u(t)$ and target states $f(x)$ are the following ones:

$$H = L_p(0,T), \quad F = L_p(0,l) \times W_p^{-1}(0,l), \quad 1 < p < \infty. \tag{4.3}$$

Here $L_p(a,b)$ is Lebesgue space of measurable functions ϕ defined on (a,b) with integrable $|\phi|^p$ on (a,b). Space $W_p^{-1}(0,l)$ is adjoint to Sobolev space $\overset{\circ}{W}{}_q^1(0,l)$ of functions $\phi \in L_q(0,l)$ having the first derivative $\phi' \in L_q(0,l)$ and vanishing at both endpoints: $\phi(0) = \phi(l) = 0$. The numbers p and q are adjoint: $1/p + 1/q = 1$. The norms are defined as follows:

$$\begin{gathered} \|u\|^p_{L_p(0,T)} = \int_0^T |u(t)|^p\,dt, \quad \|f^1\|_{W_p^{-1}(0,l)} = \sup_{\|w\|_{\overset{\circ}{W}{}_q^1(0,l)} \le 1} \langle f^1, w\rangle, \\ \|w\|^q_{\overset{\circ}{W}{}_q^1(0,l)} = \int_0^l |w'(x)|^q\,dx, \quad \|f\|_F^p = \|f^0\|^p_{L_p(0,l)} + \|f^1\|^p_{W_p^{-1}(0,l)}. \end{gathered} \tag{4.4}$$

Let us also consider adjoint problem [15, 26]:

$$\begin{gathered} p_{tt}(t,x) = p_{xx}(t,x), \quad (t,x) \in (0,T) \times (0,l), \\ p|_{x=0} = 0, \quad p|_{x=l} = 0, \quad t \in (0,T), \\ p|_{t=T} = v^0(x), \quad p_t|_{t=T} = -v^1(x), \quad x \in (0,l). \end{gathered} \tag{4.5}$$

Analogously to [11] it can be proved that linear operator

$$\mathcal{A}^* v = p_x\,|_{x=0}, \quad \mathcal{A}^* : F^* = \overset{\circ}{W}{}^1_q(0,l) \times L_q(0,l) \to H^* = L_q(0,T), \tag{4.6}$$

is well-defined and bounded: $\mathcal{A}^* \in \mathcal{L}(F^* \to H^*)$. Then its adjoint operator

$$\mathcal{A}^{**} u = \mathcal{A} u = \left(y\,|_{t=T}, y_t\,|_{t=T}\right), \quad \mathcal{A} : H \to F,$$

is also linear and bounded: $\mathcal{A} \in \mathcal{L}(H \to F)$, so the boundary control problem (4.1), (4.2) can be reformulated as equation (1) in Banach spaces H and F. We will find its normal solution u_* with property (2).

Let us prove that all assumptions B1–B7 are fulfilled for this problem. It is well known that assumptions B1, B2 and B3 are satisfied (see [20, Section 36, p. 78], [1, Theorem 3.6, p. 61] and [20, Section 37, p. 78]). Assumption B4 will be satisfied if we take function $\phi(x) = x^{p-1}$ and define functionals

$$P(u) = \frac{1}{p}\,\|u\|^p_{L_p(0,T)}, \quad K(g) = \frac{1}{q}\,\|g\|^q_{L_q(0,T)}.$$

Both of them have continuous Fréchet derivatives:

$$\langle P'(u), \overline{u}\rangle = \int_0^T |u(t)|^{p-1} \operatorname{sgn} u(t)\overline{u}(t)\,dt, \quad \forall u, \overline{u} \in L_p(0,T),$$
$$\langle K'(g), \overline{g}\rangle = \int_0^T |g(t)|^{q-1} \operatorname{sgn} g(t)\overline{g}(t)\,dt, \quad \forall g, \overline{g} \in L_q(0,T).$$

Continuity of $K'(u)$, $P'(g)$ can be established using a partial converse of the Lebesgue dominated convergence theorem [4, Theorem 4.9, p. 94]. In order to check the assumptions B5–B7 we prove observability inequality [26]:

$$\|\mathcal{A}^* v\|_{H^*} \geq \mu \|v\|_{F^*}, \quad \forall v \in F^*. \tag{4.7}$$

Theorem 4. *Let spaces H, F and their norms be defined in (4.3), (4.4), operator $\mathcal{A}^*$ be defined in (4.6) and $T \geq 2l$. Then inequality (4.7) holds true with constant $\mu = 1$.*

P r o o f. Let us denote $g(t) = (\mathcal{A}^* v)(t) = p_x(t,0)$, $t \in (0,T)$. Then fixing some $x \in (0,l)$ and integrating differential equation from (4.5) along characteristic $\{(\tau,\xi) \mid \xi \in [0,x],\ \tau = T - (x - \xi)\}$ we get

$$p_t(T,x) - p_x(T,x) = p_t(T-x,0) - p_x(T-x,0) = -g(T-x).$$

Analogously after integrating differential equation along characteristics $\tau = T - (\xi - x)$, $\xi \in [x,l]$, and $\tau = T - (l-x) - (l-\xi)$, $\xi \in [0,l]$, we obtain

$$\begin{gathered} p_t(T,x) + p_x(T,x) = p_t(T-(l-x),l) + p_x(T-(l-x),l) = p_x(T-(l-x),l), \\ -p_x(T-(l-x),l) = p_t(T-(l-x),l) - p_x(T-(l-x),l) = \\ = p_t(T-(l-x)-l,0) - p_x(T-(l-x)-l,0) = -g(T-2l+x), \end{gathered}$$

so

$$p_t(T,x) + p_x(T,x) = g(T-2l+x),$$
$$p_t(T,x) = \frac{1}{2}\left(g(T-2l+x) - g(T-x)\right), \quad p_x(T,x) = \frac{1}{2}\left(g(T-2l+x) + g(T-x)\right).$$

Then using Jensen's inequality we obtain

$$\|v\|^q_{F^*} = \int_0^l (|p_t(T,x)|^q + |p_x(T,x)|^q)\,dx \le \int_0^l (|g(T-2l+x)|^q + |g(T-x)|^q)\,dx =$$
$$= \int_{T-2l}^T |g(t)|^q\,dt \le \|g\|^q_{L_q(0,T)} = \|\mathcal{A}^* v\|^q_{H^*}.$$

It means that the constant μ in (4.7) is equal to 1. □

Remark 4. The value of $\mu = 1$ is adequate for T being close to $2l$, but becomes too rough for sufficiently large T. Using a slightly modified technique, one can obtain for μ another expression of the form $\mu = C \cdot (T-2l)$ (with a constant $C > 0$ independent on T) being more preferable for sufficiently large T.

It follows from observability inequality (4.7) that $R(\mathcal{A}) = F$ [14, Theorem 3.6, p. 13], so the assumption B5 is fulfilled. Then the closedness of $R(\mathcal{A})$ implies closedness of $R(\mathcal{A}^*)$ [14, Theorem 3.7, p. 13], and with the help of corollary 1 the validity of the assumption B6 is proved.

Remark 5. Note that, despite of closedness of $R(\mathcal{A}^*)$, even in the case of Hilbert spaces ($p = 2$) the problem (4.1), (4.2) is unstable when approximate operators $\mathcal{A}_n$ are constructed using finite difference space semi-discrete scheme, as it was shown in [26]. Using fully discrete schemes with inequal time and space mesh steps is also noted in [26] as a practice that leads to instabilities. Indirectly it was illustrated by non-regularized computations in [6].

To find a value r_* for the source norm estimate from assumption B7 take into account, that element v_* is the unique source (due to (4.7)) for the solution u_* and satisfies the following conditions:

$$\begin{aligned}\|\mathcal{A}^* v_*\|^q_{L_q(0,T)} &= \int_0^T |(\mathcal{A}^* v_*)(t)|^q\,dt = \int_0^T |(\mathcal{A}^* v_*)(t)|^{q-1} (\mathcal{A}^* v_*)(t) \operatorname{sgn}(\mathcal{A}^* v_*)(t)\,dt = \\ &= \langle \mathcal{A}^* v_*, K'(\mathcal{A}^* v_*)\rangle = \langle v_*, \mathcal{A}\mathcal{J}_H \mathcal{A}^* v_*\rangle = \langle v_*, f\rangle \le \|v_*\|_{F^*}\|f\|_F.\end{aligned} \tag{4.8}$$

Inequality (4.7) brings us to

$$\mu^q \|v_*\|^q_{F^*} \le \|\mathcal{A}^* v_*\|^q_{H^*} \le \|v_*\|_{F^*}\|f\|_F,$$

i. e.

$$\|v_*\|_{F^*} \le \mu^{-p}\|f\|_F^{p/q} \equiv r_*.$$

In our case $\mu = 1$, so $r_* = \|f\|^{p/q}$, and assumption B7 is true. In practice, if we know only approximate target f_n, we can take $r_* = \|f_n\|^{p/q} + \gamma$ with some fixed $\gamma > 0$.

Remark 6. Note that inequality of type (4.8) can be obtained not only in the case $H = L_p(0,T)$. For abstract spaces, using Asplund's theorem [5, Theorem 4.4, p. 26], the following estimate can be established for the source v_* of the solution u_*:

$$h\left(\|\mathcal{A}^* v_*\|_{H^*}\right) \le \|v_*\|_{F^*}\|f\|_F\,, \quad h(x) = x\,\phi^{-1}(x).$$

Remark 7. The method can also be applied to solve boundary control problems of type (4.1) for the one-dimensional wave equation with variable coefficients $\rho, k \in BV[0,l]$, $q \in C[0,l]$:

$$\rho(x) y_{tt}(t,x) = (k(x) y_x(t,x))_x - q(x) y(t,x), \quad (t,x) \in]0,T[\times]0,l[$$

for all $T \ge 2\int_0^l \sqrt{\rho(x)/k(x)}\,dx$. In this paper the case $\rho(x) = k(x) = 1$, $q(x) = 0$ was considered for simplicity of proving observability inequality (4.7).

5. Numerical Experiments

Numerical experiments were produced for the problem (4.1) with $l = 1$, $T = 3 = 3l > 2l$, $p = 3$ and $\mu = 1$. As a terminal target state $f = (f^0(x), f^1(x))$ we choose

$$\begin{aligned} f^0(x) &= u_*(3l - x) - u_*(l + x) + u_*(l - x), \quad 0 < x < l, \\ f^1(x) &= u_*'(3l - x) - u_*'(l + x) + u_*'(l - x), \quad 0 < x < l, \end{aligned} \tag{5.1}$$

where

$$u_*(t) = \begin{cases} 3l/4 - |t - 3l/4|\,, & 0 < t < 3l/2, \\ 3\sqrt{3}\,(|t - 7l/4| - l/4)\,, & 3l/2 < t < 2l, \\ 3l/4 - |t - 11l/4|\,, & 2l < t < 3l. \end{cases}$$

Note that at first we chose control $u_*(t)$ such that $u_* \in \mathcal{J}_H R(\mathcal{A}^*)$. After that using explicit expressions for the solution of boundary value problem (4.1) we defined target $f = \mathcal{A}u_*$, so according to corollary 1 it means that $u_*(t)$ is the solution to (2). Plots of $u_*(t)$ and $f(x)$ are shown at Figure 1 and Figure 2 respectively.

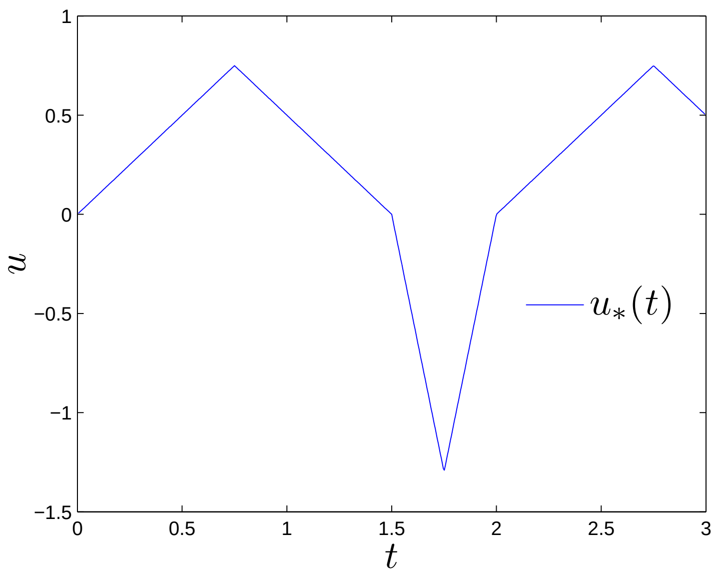
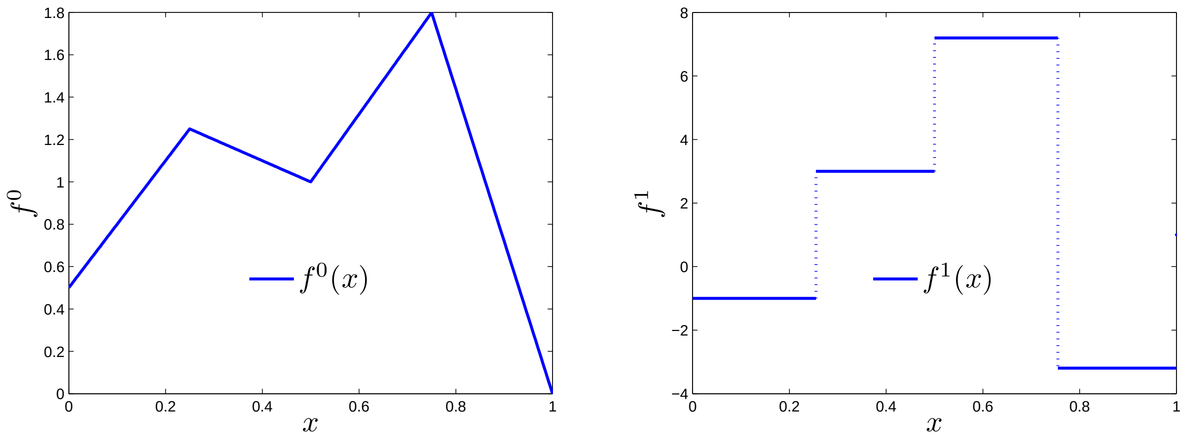


Figure 1. Plot of the exact control $u_*(t)$

Approximate operator $\mathcal{A}_n$ was built similar to [18] using three-layer explicit difference scheme on a uniform grid with M nodes on segment $[0, l]$ and N nodes on $[0, T]$. Approximate terminal state f_n was produced by discretization of functions (5.1) and by adding random noise of fixed level $\delta = \|f_n - f_d\|_F / \|f_d\|_F$, where f_d is discretized function (5.1). The Table 1 presents some relative errors $\epsilon = \|u_n - u_*\|_H / \|u_*\|_H$ (where $H = L_3(0, 1)$) of finding control $u_*(t)$ by the method, depending on grid parameters M, N and noise level δ in target state. Some typical plots of approximate controls $u_n(t)$ are presented at Figure 3 and Figure 4.

As it can be seen from Table 1, errors of the method are quite acceptable and decrease with grid refinement and noise vanishing, that agrees with the theoretical conclusions stated above. It is curious that the numerical results are sensitive to small variations in the steps of the difference grid near their equal values when the stability condition of the difference scheme is satisfied. Of course, other methods oriented to problems of the type (4.1) can give better results. The main advantages



(a) Plot of $f^0(x)$ (b) Plot of $f^1(x)$

Figure 2. Plot of the exact terminal state $f(x)$

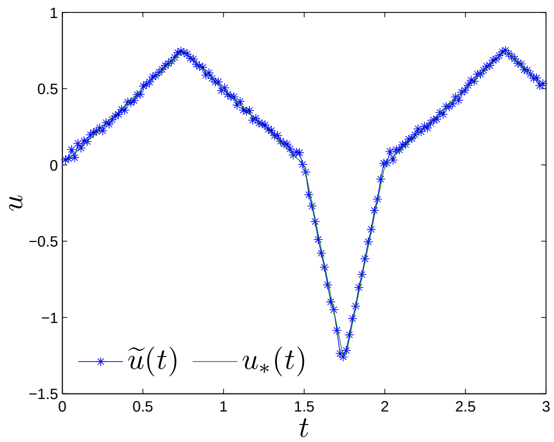
N	M	δ	ϵ
150	50	0%	6.67%
155	50	0%	4.27%
150	50	8%	20.3%
155	50	8%	19.9%
300	100	0%	3.21%
310	100	0%	2.08%
300	100	4%	19%
310	100	4%	11.5%
600	200	0%	1.16%
620	200	0%	0.97%
600	200	2%	12.6%
620	200	2%	5.11%

Table 1. Relative errors $\epsilon = \|u_n - u_*\|_H / \|u_*\|_H$ of the method

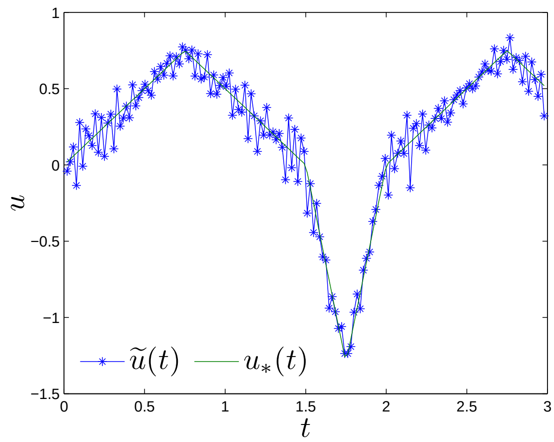
of our method are its universality, the possibility of applying to a wide class of ill-posed problems in Banach spaces and also the existence of a theoretical base in the form of assumptions B1–B7.

6. Conclusion

In the paper a numerical method for the linear equation in Banach spaces is proposed. The main advantage of the method is its applicability to problems with non-uniformly perturbed operator. However, inequalities like (4.7) with explicit values of μ can be obtained only in limited number of applications, so in practice the problem of choosing an appropriate value of the important parameter r_* can occur sufficiently difficult. We also note that for uniformly convex Banach spaces it seems possible to obtain error estimate of the proposed method.



(a) Subcase $\delta = 0\%$

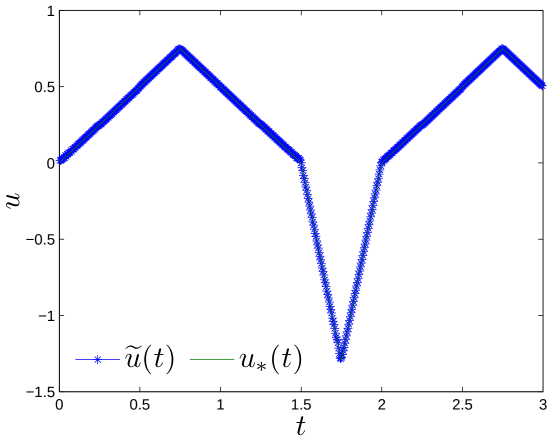


(b) Subcase $\delta = 8\%$

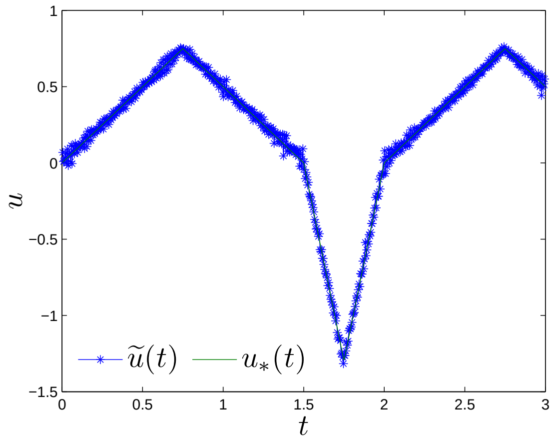
Figure 3. Plots of approximate solution $\widetilde{u}(t) = u_n(t)$ in comparison to exact solution $u_*(t)$ in the case $N = 155$, $M = 50$

REFERENCES

1. Adams R. A., Fournier J. J. F. *Sobolev Spaces.* Amsterdam: Elsevier, 2003. 320 p.
2. Bakushinskii A. B. Methods for solving monotonic variational inequalities, based on the principle of iterative regularization. *USSR Computational Mathematics and Mathematical Physics*, 1977. Vol. 17, No. 6. P. 12–24.
3. Bakushinsky A., Goncharsky A. *III-Posed Problems: Theory and Applications.* Dordrecht: Kluwer Academic Publishers, 1994. 258 p. DOI: 10.1007/978-94-011-1026-6
4. Brezis H. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* New York: Springer, 2011. 599 p. DOI: 10.1007/978-0-387-70914-7
5. Cioranescu I. *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems.* Dordrecht: Kluwer Academic Publishers, 1990. 260 p. DOI: 10.1007/978-94-009-2121-4
6. Dryazhenkov A. A., Potapov M. M. Constructive observability inequalities for weak generalized solutions of the wave equation with elastic restraint. *Comput. Math. Math. Phys.*, 2014. Vol. 54, No. 6. P. 939–952. DOI: 10.1134/S0965542514060062
7. Dunford N., Schwartz J. T. *Linear Operators. Part I: General Theory.* New York: Interscience Publishers, 1958. 872 p.
8. Ekeland I., Temam R. *Convex Analysis and Variational Problems.* Amsterdam: North-Holland Publishing Company, 1976. 394 p. DOI: 10.1137/1.9781611971088
9. Ekeland I., Turnbull T. *Infinite-Dimensional Optimization and Convexity.* Chicago: The University of Chicago Press, 1983. 174 p.
10. Engl H. W., Hanke M., Neubauer A. *Regularization of Inverse Problems.* Dordrecht: Kluwer Academic Publishers, 1996. 322 p.
11. Il'in V. A., Kuleshov A. A. On some properties of generalized solutions of the wave equation in the classes L_p and W_p^1 for $p \geq 1$. *Differ. Equ.*, 2012. Vol. 48, No. 11. P. 1470–1476. DOI: 10.1134/S0012266112110043
12. Ivanov V. K. On linear problems that are not well-posed. *Soviet Mathematics Doklady*, 1962. Vol. 3. P. 981–983.
13. Kantorovich L. V., Akilov G. P. *Functional Analysis.* Oxford: Pergamon Press, 1982. 604 p. DOI: 10.1016/C2013-0-03044-7
14. Krein S. G. *Linear Equations in Banach Spaces.* Boston: Birkhäuser, 1982. 106 p. DOI: 10.1007/978-1-4684-8068-9
15. Lions J.-L. Exact controllability, stabilization and perturbations for distributed systems. *SIAM Rev.*, 1988. Vol. 30, No. 1. P. 1–68. DOI: 10.1137/1030001



(a) Subcase $\delta = 0\%$



(b) Subcase $\delta = 2\%$

Figure 4. Plots of approximate solution $\widetilde{u}(t) = u_n(t)$ in comparison to exact solution $u_*(t)$ in the case $N = 620$, $M = 200$

16. Morozov V. A. Regularization of incorrectly posed problems and the choice of regularization parameter. *USSR Computational Mathematics and Mathematical Physics*, 1966. Vol. 6, No. 1. P. 242–251. DOI: 10.1016/0041-5553(66)90046-2
17. Phillips D. L. A technique for the numerical solution of certain integral equations of the first kind. *J. ACM*, 1962. Vol. 9, No. 1. P. 84–97. DOI: 10.1145/321105.321114
18. Potapov M. M. Strong convergence of difference approximations for problems of boundary control and observation for the wave equation. *Comput. Math. Math. Phys.*, 1998. Vol. 38, No. 3. P. 373–383.
19. Potapov M. M. A stable method for solving linear equations with nonuniformly perturbed operators. *Dokl. Math.*, 1999. Vol. 59, No. 2. P. 286–288.
20. Riesz F., Sz.-Nagy B. *Functional Analysis.* London: Blackie & Son Limited, 1956. 468 p.
21. Scherzer O., Grasmair M., Grossauer H., Haltmeier M., Lenzen F. *Variational Methods in Imaging.* New York: Springer, 2009. 320 p. DOI: 10.1007/978-0-387-69277-7
22. Schuster T., Kaltenbacher B., Hofmann B., Kazimierski K. S. *Regularization Methods in Banach Spaces.* Berlin: De Gruyter, 2012. 283 p.
23. Tikhonov A. N. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 1963. Vol. 4, No. 4. P. 1035–1038.
24. Tikhonov A. N., Arsenin V. Y. *Solution of Ill-posed Problems.* Washington: Winston & Sons, 1977. 258 p.
25. Tikhonov A. N., Leonov A. S., Yagola A. G. *Nonlinear Ill-posed Problems.* London: Chapman & Hall, 1998. 386 p.
26. Zuazua E. Propagation, observation, and control of waves approximated by finite difference methods. *SIAM Rev.*, 2005. Vol. 47, No. 2. P. 197–243. DOI: 10.1137/S0036144503432862

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 69–78

DOI: 10.15826/umj.2018.2.008

AUTOMORPHISMS OF A DISTANCE-REGULAR GRAPH WITH INTERSECTION ARRAY $\{39, 36, 4; 1, 1, 36\}$[1].

Konstantin S. Efimov

Ural State University of Economics,
62 March 8th Str., Ekaterinburg, Russia, 620144

konstantin.s.efimov@gmail.com

Alexander A. Makhnev

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, Russia, 620990

makhnev@imm.uran.ru

Abstract: Makhnev and Nirova have found intersection arrays of distance-regular graphs with no more than 4096 vertices, in which $\lambda = 2$ and $\mu = 1$. They proposed the program of investigation of distance-regular graphs with $\lambda = 2$ and $\mu = 1$. In this paper the automorphisms of a distance-regular graph with intersection array $\{39, 36, 4; 1, 1, 36\}$ are studied.

Keywords: Strongly regular graph, Distance-regular graph.

Introduction

We consider undirected graphs without loops and multiple edges. Our terminology and notation are mostly standard and could be found in [1]. Given a vertex a in a graph Γ, we denote by $\Gamma_i(a)$ the subgraph induced by Γ on the set of all the vertices of Γ, that are at the distance i from a. The subgraph $[a] = \Gamma_1(a)$ is called the *neighbourhood of a vertex* a. Let $\Gamma(a) = \Gamma_1(a)$, $a^\perp = \{a\} \cup \Gamma(a)$. If graph Γ is fixed, then we write $[a]$ instead of $\Gamma(a)$.

The incidence system with the set of points P and the set of lines $\mathcal{L}$ is called α*-partial geometry of order* (s, t) if each line contains exactly $s + 1$ points, each point lies exactly on $t + 1$ lines, any two points lie on no more than one line, and for any antiflag $(a, l) \in (P, \mathcal{L})$ there are exactly α lines passing through a and intersecting l. This geometry is denoted by $pG_\alpha(s, t)$.

In the case $\alpha = 1$, the geometry $pG_\alpha(s, t)$ is called a *generalized quadrangle* and is denoted by $GQ(s, t)$. A point graph of this geometry is defined on the set of points P and two points are adjacent if they lie on a line. The point graph of a geometry $pG_\alpha(s, t)$ is strongly regular with parameters $v = (s + 1)(1 + st/\alpha)$, $k = s(t + 1)$, $\lambda = s - 1 + t(\alpha - 1)$, $\mu = \alpha(t + 1)$. A strongly regular graph with such parameters for some natural numbers α, s, t is called a *pseudo-geometric graph* for $pG_\alpha(s, t)$.

If vertices u, w are at distance i in Γ, then by $b_i(u, w)$ (respectively, $c_i(u, w)$) we denote the number of vertices in $\Gamma_{i+1}(u) \cap [w]$ (respectively, $\Gamma_{i-1}(u) \cap [w]$). A graph Γ of diameter d is called *distance-regular with intersection array* $\{b_0, b_1, \ldots, b_{d-1}; c_1, \ldots, c_d\}$ if the values $b_i(u, w)$ and $c_i(u, w)$ do not depend on the choice of vertices u, w at distance i in Γ for each $i = 0, ..., d$. Note that, for a

[1]This work is partially supported by RSF, project 14-11-00061P

distance-regular graph, b_0 is the degree of the graph and $c_1 = 1$. For a subset X of automorphisms of a graph Γ, $\mathrm{Fix}(X)$ denotes the set of all vertices of Γ, fixed with respect to any automorphism of X. Further, by $p^l_{ij}(x, y)$ we denote the number of vertices in a subgraph $\Gamma_i(x) \cap \Gamma_j(y)$ for vertices x, y at distance l in Γ.

A graph is said to be vertex-symmetric if its automorphism group acts transitively on the set of its vertices.

In [2], intersection arrays of distance-regular graphs with $\lambda = 2$, $\mu = 1$ and with the number of vertices at most 4096 were found. A.A. Makhnev and M.S. Nirova proposed an investigation program of automorphisms of distance-regular graphs from the obtained list.

Proposition 1. [2] *Let Γ be a distance-regular graph with $\lambda = 2$, $\mu = 1$, which has at most 4096 vertices. Then Γ has one of the following intersection arrays:*

(1) $\{21, 18; 1, 1\}(v = 400)$;

(2) $\{6, 3, 3, 3; 1, 1, 1, 2\}$ (Γ *is a generalized octagon of order* $(3, 1)$, $v = 160$), $\{6, 3, 3; 1, 1, 2\}$ (Γ *is a generalized hexagon of order* $(3, 1)$, $v = 52$), $\{12, 9, 9; 1, 1, 4\}$ (Γ *is a generalized hexagon of order* $(3, 3)$, $v = 364$), $\{6, 3, 3, 3, 3, 3; 1, 1, 1, 1, 1, 2\}$ (Γ *is a generalized dodecagon of order* $(3, 1)$, $v = 1456$);

(3) $\{18, 15, 9; 1, 1, 10\}(v = 1 + 18 + 270 + 243 = 532$, Γ_3 *is a strongly regular graph*); $\{33, 30, 8; 1, 1, 30\}$, $\{39, 36, 4; 1, 1, 36\}$, $\{21, 18, 12, 4; 1, 1, 6, 21\}$.

In this paper we study automorphisms of a hypothetical distance-regular graph Γ with intersection array $\{39, 36, 4; 1, 1, 36\}$. The maximal order of a clique C in Γ is not more than 4. A graph with intersection array $\{39, 36, 4; 1, 1, 36\}$ has $v = 1 + 39 + 1404 + 156 = 1600$ vertices and the spectrum $39^1, 7^{675}, -1^{156}, -6^{768}$.

Theorem 1. *Let Γ be a distance-regular graph with intersection array $\{39, 36, 4; 1, 1, 36\}$, $G = \mathrm{Aut}(\Gamma)$, g is an element of prime order p in G and $\Omega = \mathrm{Fix}(g)$ contains exactly s vertices in t antipodal classes. Then $\pi(G) \subseteq \{2, 3, 5\}$ and one of the following statements holds*:

(1) *Ω is an empty graph and either $p = 2$, $\alpha_1(g) = 10r + 26m + 12$ and $\alpha_3(g) = 80r$ or $p = 5$, $\alpha_1(g) = 65n + 10l + 10$ and $\alpha_3(g) = 200l$;*

(2) *Ω is an n-clique and one of the following statements holds*:

(*i*) *$n = 1$, $p = 3$, $\alpha_1(g) = 15l + 24 + 39m$ and $\alpha_3(g) = 120l + 36$,*

(*ii*) *$n = 2$, $p = 2$, $\alpha_1(g) = 10l + 26m$ and $\alpha_3(g) = 80l - 8$,*

(*iii*) *$n = 4$, $p = 2$, $\alpha_1(g) = 10l + 26m + 14$ and $\alpha_3(g) = 80l - 16$ or $p = 3$, $\alpha_1(g) = 10l + 39m + 1$, l is congruent to -1 modulo 3 and $\alpha_3(g) = 120l + 24$;*

(3) *Ω consists of n vertices pairwise at distance 3 in Γ, $p = 3$, $n \in \{4, 7, ..., 40\}$, $\alpha_3(g) = 120l + 40 - 4n$ and $\alpha_1(g) = 15l + 30 + 39m - 6n$;*

(4) *Ω contains an edge and is a union of isolated cliques, any two vertices of different cliques are at distance 3 in Γ, and either $p = 3$ and the orders of these cliques are 1 or 4, or $p = 2$ and the orders of these cliques are 2 or 4;*

(5) *Ω contains vertices that are at distance 2 in Γ and $p \leq 3$.*

If Γ is a distance-regular graph with the intersection array $\{39, 36, 4; 1, 1, 36\}$ then Γ_3 is a pseudo-geometric for $pG_3(39, 3)$.

Theorem 2. *Let Γ be a strongly regular graph with parameters $(1600, 156, 44, 12)$, $G = \mathrm{Aut}(\Gamma)$, g is an element of prime order p in G and $\Delta = \mathrm{Fix}(g)$. Then $p \leq 43$ and the following statements hold*:

(1) *if Δ is an empty graph, then $p = 2$ and $\alpha_1(g) = 80s$ or $p = 5$ and $\alpha_1(g) = 200t$;*

(2) *if Δ is an n-clique, then one of the following statements holds*:

(i) $n = 1$, $p = 2$ *and* $\alpha_1(g) = 80s - 4$, *or* $p = 3$ *and* $\alpha_1(g) = 120t + 36$, *or* $p = 13$ *and* $\alpha_1(g) = 520l + 156$,

(ii) $n \in \{4, 7, 10, ..., 40\}$, $p = 3$ *and* $\alpha_1(g) = 120t + 40 - 4n$,

(iii) $n = 9$, $p = 37$ *and* $\alpha_1(g) = 444$;

(3) *if* Δ *is an* m*-coclique, where* $m > 1$, *then either* $p = 2$, $m \in \{4, 6, 8, ..., 40\}$ *and* $\alpha_1(g) = 80s - 4m$ *or* $p = 3$, $m \in \{4, 7, 10, ..., 40\}$ *and* $\alpha_1(g) = 120t + 40 - 4m$;

(4) *if* Δ *contains an edge and is an union of isolated cliques, then* $p = 3$;

(5) *if* Δ *contains a geodesic* 2*-path, then* $p \leq 43$.

Corollary 1. *Let* Γ *be a distance-regular graph with intersection array* $\{39, 36, 4; 1, 1, 36\}$ *and nonsolvable group* $G = \mathrm{Aut}(\Gamma)$ *acts transitively on the set of vertices of* Γ. *If* a *is a vertex of* Γ, $\bar{T}$ *is the socle of the group* $\bar{G} = G/O_{5'}(G)$, *then* $\bar{T} = L \times M$, *and each of subgroups* L, M *is isomorphic to one of the following groups:* Z_5, A_5, A_6 *or* $PSp(4, 3)$.

If $|\bar{T} : \bar{T}_a| = 40^2$, then $O_{5'}(G) = 1$ and this case is realized if one of the following statements holds:

(1) $L \cong M \cong PSp(4, 3)$, $|L : L_a| = |M : M_a| = 40$,

(2) $L \cong PSp(4, 3)$, $|L : L_a| = 40$, $M \cong A_6$ and $|M_a| = 9$,

(3) $L \cong M \cong A_6$ and $|L_a| = |M_a| = 9$.

1. Proof of Theorem 2

First we give auxiliary results.

Lemma 1. [2, Theorem 3.2] *Let* Γ *be a strongly regular graph with parameters* (v, k, λ, μ) *and with the second eigenvalue* r. *If* g *is an automorphism of* Γ *and* $\Delta = \mathrm{Fix}(g)$, *then*

$$|\Delta| \leq v \cdot \max\{\lambda, \mu\}/(k - r).$$

By Lemma 1, for a strongly regular graph with parameters $(1600, 156, 44, 12)$ we have $|\Delta| \leq 1600 \cdot \max\{44, 12\}/(156 - 36)$, $|\Delta| \leq 586$.

Lemma 2. *Let* Γ *be a distance regular graph with intersection array* $\{39, 36, 4; 1, 1, 36\}$. *Then for intersection numbers of* Γ *the following statements hold*:

(1) $p^1_{11} = 2$, $p^1_{12} = 36$, $p^1_{22} = 1224$, $p^1_{23} = 144$, $p^1_{33} = 12$;

(2) $p^2_{11} = 1$, $p^2_{12} = 34$, $p^2_{13} = 4$, $p^2_{22} = 1229$, $p^2_{23} = 140$, $p^2_{33} = 12$;

(3) $p^3_{12} = 36$, $p^3_{13} = 3$, $p^3_{22} = 1260$, $p^3_{23} = 108$, $p^3_{33} = 44$.

P r o o f. This follows from [1, Lemma 4.1.7]. □

The proofs of Theorems 1 and 2 are based on Higman's method of working with automorphisms of a distance-regular graph, presented in the third chapter of Cameron's book [4].

Let Γ be a distance-regular graph of diameter d with v vertices. Then we have a symmetric association scheme $(X, \mathcal{R})$ with d classes, where X is the set of vertices of Γ and $R_i = \{(u, w) \in X^2 | d(u, w) = i\}$. For a vertex $u \in X$ we set $k_i = |\Gamma_i(u)|$. Let A_i be an adjacency matrix of graph Γ_i. Then $A_i A_j = \sum p^l_{ij} A_l$ for some integer numbers $p^l_{ij} \geq 0$, which are called the intersection numbers. Note that $p^l_{ij} = |\Gamma_i(u) \cap \Gamma_j(w)|$ for any vertices u, w with $d(u, w) = l$.

Let P_i be a matrix in which in the (j, l)-th entry is p^l_{ij}. Then eigenvalues $k = p_1(0), ..., p_1(d)$ of the matrix P_1 are eigenvalues of Γ with multiplicities $m_0 = 1, ..., m_d$, respectively. The matrices P

and Q with $P_{ij} = p_j(i)$ and $Q_{ji} = m_j p_i(j)/k_i$ are called the first and the second eigenmatrices of Γ, respectively, and $PQ = QP = vI$, where I is an identity matrix of order $d+1$.

The permutation representation of the group $G = \mathrm{Aut}(\Gamma)$ on the vertex set of Γ naturally gives the monomial matrix representation ψ of a group G in $GL(v, \mathbb{C})$. The space $\mathbb{C}^v$ is an orthogonal direct sum of the eigenspaces $W_0, W_1, ..., W_d$ of the adjacent matrix $A = A_1$ of Γ. For every $g \in G$, we have $\psi(g)A = A\psi(g)$, so each subspace W_i is $\psi(G)$-invariant. Let χ_i be the character of a representation ψ_{W_i}. Then for $g \in G$ we obtain $\chi_i(g) = v^{-1}\sum_{j=0}^{d} Q_{ij}\alpha_j(g)$, where $\alpha_j(g)$ is the number of vertices x of X such that $d(x, x^g) = j$.

Lemma 3. *Let Γ be a strongly regular graph with parameters $(1600, 156, 44, 12)$ and with the spectrum $156^1, 36^{156}, -4^{1443}$, $G = \mathrm{Aut}(\Gamma)$. If $g \in G$, χ_1 is the character of ψ_{W_1}, where $\dim(W_1) = 156$, then $\alpha_i(g) = \alpha_i(g^l)$ for any natural number l, coprime to $|g|$, $\chi_1(g) = (4\alpha_0(g) + \alpha_1(g))/40 - 4$. Moreover, if $|g| = p$ is a prime, then $\chi_1(g) - 156$ is divisible by p.*

P r o o f. We have

$$Q = \begin{pmatrix} 1 & 1 & 1 \\ 156 & 36 & -4 \\ 1443 & -37 & 3 \end{pmatrix}.$$

So, $\chi_1(g) = (39\alpha_0(g) + 9\alpha_1(g) - \alpha_2(g))/400$. Note that $\alpha_2(g) = 1600 - \alpha_0(g) - \alpha_1(g)$, so $\chi_1(g) = (4\alpha_0(g) + \alpha_1(g))/40 - 4$. The remaining statements of the lemma follow from Lemma 2 [5]. $\square$

Lemma 4. *Let Γ be a distance-regular graph with intersection array $\{39, 36, 4; 1, 1, 36\}$, $G = \mathrm{Aut}(\Gamma)$. If $g \in G$, χ_1 is the character of ψ_{W_1}, where $\dim(W_1) = 675$, χ_2 is the character of ψ_{W_2}, where $\dim(W_2) = 156$, then $\alpha_i(g) = \alpha_i(g^l)$ for any natural number l coprime to $|g|$, $\chi_1(g) = (44\alpha_0(g) + 8\alpha_1(g) - \alpha_3(g))/104 - 25/13$ and $\chi_2(g) = (4\alpha_0(g) + \alpha_3(g))/40 - 4$. Moreover, if $|g| = p$ is a prime, then $\chi_1(g) - 675$ and $\chi_2(g) - 156$ are divisible by p.*

P r o o f. We have

$$Q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 675 & 1575/13 & -25/13 & -225/13 \\ 156 & -4 & -4 & 36 \\ 768 & -1536/13 & 64/13 & -256/13 \end{pmatrix}.$$

This means $\chi_1(g) = (351\alpha_0(g) + 63\alpha_1(g) - \alpha_2(g) - 9\alpha_3(g))/832$. Note that $\alpha_2(g) = 1600 - \alpha_0(g) - \alpha_1(g) - \alpha_3(g)$, so $\chi_1(g) = (44\alpha_0(g) + 8\alpha_1(g) - \alpha_3(g))/104 - 25/13$.

Similarly, $\chi_2(g) = (39\alpha_0(g) - \alpha_1(g) - \alpha_2(g) + 9\alpha_3(g))/400$. Note that $\alpha_1(g) + \alpha_2(g) = 1600 - \alpha_0(g) - \alpha_3(g)$, so $\chi_2(g) = (4\alpha_0(g) + \alpha_3(g))/40 - 4$.

The remaining statements of this lemma follow from Lemma 2 of [5]. $\square$.

In Lemmas 5–7 we suppose that Γ is a strongly regular graph with parameters $(1600, 156, 44, 12)$, $G = \mathrm{Aut}(\Gamma)$, g is an element of prime order p from G, $\alpha_i(g) = pw_i$ for $i > 0$ and $\Delta = \mathrm{Fix}(g)$. By Delsarts's boundary the maximal order of a clique K in Γ is not greater than $1 - k/\theta_d$, so $|K| \le 40$. Due to Hoffman's boundary the maximum order of a coclique C in Γ is not greater than $-v\theta_d/(k - \theta_d)$, so $|C| \le 40$.

Lemma 5. *The following statements hold*:

(1) *if Δ is an empty graph, then either $p = 2$ and $\alpha_1(g) = 80s$ or $p = 5$ and $\alpha_1(g) = 200t$;*

(2) *if Δ is an n-clique, then one of the following statements holds*:

(i) $n = 1$, $p = 2$ *and* $\alpha_1(g) = 80s - 4$, *or* $p = 3$ *and* $\alpha_1(g) = 120t + 36$, *or* $p = 13$ *and* $\alpha_1(g) = 520l + 156$,

(ii) $n \in \{4, 7, 10, ..., 40\}$, $p = 3$ *and* $\alpha_1(g) = 120t + 40 - 4n$,

(iii) $n = 9$, $p = 37$ *and* $\alpha_1(g) = 444$;

(3) *if* Δ *is an* m*-coclique, where* $m > 1$*, then* $p = 2$, $m \in \{4, 6, 8, ..., 40\}$ *and* $\alpha_1(g) = 80s - 4m$ *or* $p = 3$, $m \in \{4, 7, 10, ..., 40\}$ *and* $\alpha_1(g) = 120t + 40 - 4m$;

(4) *if* Δ *contains an edge and is an union of isolated cliques, then* $p = 3$.

P r o o f. Let Δ be an empty graph. As $v = 2^6 \cdot 25$, then p is equal to 2 or 5.

In the case $p = 2$ we have $\chi_1(g) = \alpha_1(g)/40 - 4$ and $\alpha_1(g) = 80s$.

In the case $p = 5$ we have $\chi_1(g) = \alpha_1(g)/40 - 4$ and $\alpha_1(g) = 200t$.

Let Δ be an n-clique. If $n = 1$, then p divides 156 and 1443, therefore $p \in \{2, 3, 13\}$. In the case $p = 2$ we have $\chi_1(g) = (4 + \alpha_1(g))/40 - 4$ and $\alpha_1(g) = 80s - 4$.

In the case $p = 3$ we have $\chi_1(g) = (4 + \alpha_1(g))/40 - 4$ and the number $(4 + 3w_1)/40$ is congruent to 1 modulo 3. Hence, $4 + 3w_1 = 120t + 40$ and $\alpha_1(g) = 120t + 36$.

In the case $p = 13$ we have $\chi_1(g) = (4 + 13w_1)/40 - 4$ and the number $(4 + 13w_1)/40$ is congruent to 4 modulo 13. Hence, $4 + 13w_1 = 520l + 160$ and $\alpha_1(g) = 520l + 156$.

If $n > 1$, then for any two vertices $a, b \in \Delta$ the element g acts without fixed points on $[a] \cap [b] - \Delta$, on $[a] - b^\perp$ and on $\Gamma - (a^\perp \cup b^\perp)$. Hence, p divides $46 - n$, 111 and 1332, therefore $p \in \{3, 37\}$.

In the case $p = 3$ we have $n \in \{4, 7, 10, ..., 40\}$. Further, $\chi_1(g) = (4n + \alpha_1(g))/40 - 4$ and the number $(4n + \alpha_1(g))/40$ is congruent to 1 modulo 3. Hence, $4n + 3w_1 = 120t + 40$ and $\alpha_1(g) = 120t + 40 - 4n$.

In the case $p = 37$ we have $n = 9$. Further, $\chi_1(g) = (36 + \alpha_1(g))/40 - 4$ and the number $(36 + \alpha_1(g))/40$ is congruent to 12 modulo 37. Hence, $\alpha_1(g) = 444$.

Let Δ be an m-coclique, where $m > 1$. Then for any two vertices $a, b \in \Delta$ the element g acts without fixed points on $[a] \cap [b]$, on $[a] - b^\perp$ and on $\Gamma - (a^\perp \cup b^\perp \cup \Delta)$. Hence, p divides 12, 144 and $1300 - m$, therefore $p \in \{2, 3\}$.

In the case $p = 2$ we have $m \in \{4, 6, 8, ..., 40\}$. Further, $\chi_1(g) = (4m + \alpha_1(g))/40 - 4$ and the number $(4m + \alpha_1(g))/40$ is even. Hence, $\alpha_1(g) = 80s - 4m$.

In the case $p = 3$ we have $m \in \{4, 7, 10, ..., 40\}$. Further, $\chi_1(g) = (4m + \alpha_1(g))/40 - 4$ and the number $(4m + \alpha_1(g))/40$ is congruent to 1 modulo 3. Hence, $\alpha_1(g) = 120t + 40 - 4m$.

Let Δ contains an edge and is a union of isolated cliques. Then p divides 12 and 111, therefore $p = 3$. □

Lemma 6. *If* $[a] \subset \Delta$ *for some vertex* a*, then for any vertex* $u \in \Gamma_2(a) - \Delta$ *the orbit of* $u^{\langle g \rangle}$ *is a clique or a coclique, and one of the following statements holds:*

(1) *if on* $\Gamma - \Delta$ *there are no coclique orbits, then* $\alpha_1(g) = 1600 - \alpha_0(g)$, $\alpha_0(g) = 40l$ *and either*

(i) $l = 4$, $p = 2, 3$ *or*

(ii) $l = 5$, $p = 5, 7$, *or*

(iii) $l = 6$, $p = 2$, *or*

(iv) $l = 7$, $p = 3, 11$, *or*

(v) $l = 8$, $p = 2$, *or*

(vi) $l = 10$, $p = 2, 3, 5$, *or*

(vii) $l = 12$, $p = 2, 7$, *or*

(viii) $l = 13$, $p = 3$, *or*

(ix) $l = 14$, $p = 2$;

(2) *if on* $\Gamma - \Delta$ *there is a coclique orbit, then* $p \leq 3$*, and if* $a^\perp = \Delta$*, then* $p = 3$ *and* $\alpha_1(g) = 120l + 12$.

P r o o f. Let $[a] \subset \Delta$ for some vertex a. Then for any vertex $u \in \Gamma_2(a) - \Delta$ the orbit $u^{\langle g\rangle}$ doesn't contain a geodesic 2-pathes and is a clique or a coclique.

In the case $p \geq 13$ a subgraph $[a] \cap [u]$ is a 12-clique and for two vertices $b, c \in [a] \cap [u]$ a subgraph $[b] \cap [c]$ contains a, 10 vertices from $[a] \cap [u]$ and p vertices from $u^{\langle g\rangle}$, so $11 + p \leq 44$, therefore $p \leq 31$.

If on $\Gamma - \Delta$ there are no coclique orbits, then $\alpha_1(g) = v - |\Delta|$ and for a vertex $u' \in u^{\langle g\rangle} - \{u\}$ a subgraph $[u] \cap [u']$ contains $p - 2$ vertices from $u^{\langle g\rangle}$ and 12 vertices from Δ. Further, $\chi_1(g) = (3\alpha_0(g) + 1600)/40 - 4$, $\chi_1(g) - 156$ is divisible by p and p divides $3\alpha_0(g)/40 - 120$. We denote $\alpha_0(g) = 40l$. Then $4 \leq l \leq 14$, p divides $40(40 - l)$ and $3(40 - l)$. Thus, either $l = 4$, $p = 2, 3$, or $l = 5$, $p = 5, 7$, or $l = 6$, $p = 2, 17$, or $l = 7$, $p = 3, 11$, or $l = 8$, $p = 2$, or $l = 9$, $p = 31$, or $l = 10$, $p = 2, 3, 5$, or $l = 11$, $p = 29$, or $l = 12$, $p = 2, 7$, or $l = 13$, $p = 3$, or $l = 14$, $p = 2, 13$. In the case $p \geq 13$ a subgraph $[a] \cap [u]$ is a 12-clique and $p \leq 23$.

Let $p = 17$ and $b \in \Delta - a^\perp$. Then $|\Delta(b) - a^\perp| \leq 82$ and $|[b] - \Delta| \geq 68$. For $w \in [b] - \Delta$ we have $[a] \cap [w] = [a] \cap [b]$ (otherwise $w^{\langle g\rangle}$ is contained in $[b] \cap [c]$ for $c \in [a] \cap [w] - [b]$). A contradiction with a fact that for two vertices $c, d \in [a] \cap [w]$ a subgraph $[c] \cap [d]$ contains 68 vertices from $[b] - \Delta$.

Let $p = 13$. Then $|\Delta - a^\perp| = 403$. If $b \in \Delta - a^\perp$ and $|[b] - \Delta| = 13$, then for any $w \in [b] - \Delta$ we have $[a] \cap [w] = [a] \cap [b]$ (otherwise $w^{\langle g\rangle}$ is contained in $[b] \cap [c]$ for a vertex $c \in [a] \cap [w] - [b]$). Further, $[b] \cap [w]$ contains 12 vertices from $w^{\langle g\rangle}$ and 32 vertices from $\Delta(b)$. Hence, for $w' \in w^{\langle g\rangle} - \{w\}$ a subgraph $[w] \cap [w']$ contains b, 32-clique from $\Delta(b)$ and 11 vertices from $w^{\langle g\rangle}$. A contradiction with a fact that the order of a clique in Γ is not greater than 40.

If $b \in \Delta - a^\perp$ and $|[b] - \Delta| = 26$, then $[b] - \Delta = u^{\langle g\rangle} \cup w^{\langle g\rangle}$. As above, $[a] \cap [u] = [a] \cap [w] = [a] \cap [b]$, therefore a subgraph $\{b\} \cup ([a] \cap [b]) \cup u^{\langle g\rangle} \cup w^{\langle g\rangle}$ is a 39-clique. If $e \in [u] \cap \Delta(b) - [w]$, then $[e] \cap [w]$ contains 13 vertices from $u^{\langle g\rangle}$, a contradiction. So, $\{b\} \cup ([u] \cap \Delta(b)) \cup u^{\langle g\rangle} \cup w^{\langle g\rangle}$ is a 46-coclique, a contradiction. If $b \in \Delta - a^\perp$ and $|[b] - \Delta| \geq 39$, then for any two vertices $c, d \in [a] \cap [b]$ a subgraph $[c] \cap [d]$ contains a, b and 39 vertices from $[b] - \Delta$, a contradiction. Statement (1) is proved.

Let on $\Gamma - \Delta$ there is a coclique orbit $u^{\langle g\rangle}$. Then $[u^{g_i}] \cap [u^{g_j}]$ does not intersect $\Gamma - \Delta$ for distinct vertices u^{g_i}, u^{g_j}, so $145p \leq |\Gamma - \Delta| \leq 1443$, therefore $p \leq 7$.

Let us show that $p \leq 3$.

Let $c \in [a] \cap [u]$ and $[c] \cap [u]$ contains exactly γ vertices from $[a] \cap [u]$. Then $[c] \cap [u]$ contains $44 - \gamma$ vertices outside of Δ (lying in distinct $\langle g\rangle$-orbits) and $p(44 - \gamma) \leq |[c] - \Delta| \leq 156 - 45 = 111$. Hence, $32p \leq 111$.

If $a^\perp = \Delta$, then $\alpha_0(g) = 157$, p divides 1443 and $p = 3$. Further, $\chi_1(g) = (628 + \alpha_1(g))/40 - 4$, $(628 + \alpha_1(g))/40$ is congruent to 1 modulo 3 and $\alpha_1(g) = 120l + 12$. □

Lemma 7. *The following statements hold*:

(1) Γ *does not contain proper strongly regular subgraphs with parameters* $(v', k', 44, 12)$;

(2) $p \leq 43$.

P r o o f. Assume that Γ contains proper strongly regular subgraph Σ with parameters $(v', k', 44, 12)$. Then $4(k' - 12) + 32^2 = n^2$, therefore $n = 2l$, $k' = l^2 - 244$, $l \geq 16$, Σ has nonprincipal eigenvalues $16 + l, 16 - l$ and multiplicity of $16 + l$ is equal to $(l - 17)(l^2 - 244)(l^2 + l - 260)/24l$. If l is odd, then 8 divides $(l - 17)(l^2 + l - 20)$, l divides $17 \cdot 61 \cdot 65$ and $l \in \{5, 13\}$. If l is even, then 3 divides $(l - 2)(l^2 - 1)(l^2 + l - 2)$ and $l = 16$. In all cases we have contradictions.

If $p \geq 47$, then Δ is a strongly regular graph with parameters $(v', k', 44, 12)$, so $\Delta = \Gamma$, a contradiction. □

Theorem 2 follows from Lemmas 5–7.

2. Proof of Theorem 1

In Lemmas 8–9 it is assumed that Γ is a distance-regular graph with intersection array $\{39, 36, 4; 1, 1, 36\}$, $G = \mathrm{Aut}(\Gamma)$, g is an element of prime order p from G, $\alpha_i(g) = pw_i$ for $i > 0$ and $\Omega = \mathrm{Fix}(g)$.

Lemma 8. *The following statements hold*:

(1) *if Ω is an empty graph, then either $p = 2$, $\alpha_1(g) = 10r + 26m + 12$ and $\alpha_3(g) = 80r = 1600 - \alpha_1(g)$ or $p = 5$, $\alpha_1(g) = 65n + 10l + 10$ and $\alpha_3(g) = 200l$;*

(2) *if Ω is an n-clique, then one of the following statements holds*:

(*i*) *$n = 1$, $p = 3$, $\alpha_1(g) = 15l + 24 + 39m$ and $\alpha_3(g) = 120l + 36$,*

(*ii*) *$n = 2$, $p = 2$, $\alpha_1(g) = 10l + 26m$ and $\alpha_3(g) = 80l - 8$,*

(*iii*) *or $n = 4$, $p = 2$, $\alpha_1(g) = 10l + 26m + 14$ and $\alpha_3(g) = 80l - 16$ or $p = 3$, $\alpha_1(g) = 10l + 39m + 1$, l is congruent to -1 modulo 3 and $\alpha_3(g) = 120l + 24$;*

(3) *if Ω consists of n vertices at distance 3 in Γ, then $p = 3$, $n \in \{4, 7, 10, ..., 40\}$, $\alpha_3(g) = 120l + 40 - 4n$ and $\alpha_1(g) = 15l + 30 + 39m - 6n$;*

(4) *if Ω contains an edge and doesn't contain vertices at distance 2 in Γ, then Ω is an union of isolated cliques and any two vertices from different cliques are at distance 3 in Γ, either $p = 3$ and the orders of these cliques are equal to 1 or 4, or $p = 2$ and the orders of these cliques are equal to 2 or 4.*

P r o o f. Let Ω be an empty graph and $\alpha_i(g) = pw_i$ for $i \geq 1$. As $v = 1600$, then p is equal to 2 or 5.

Let $p = 2$. Then $w_1 + w_2 + w_3 = 800$ and $\chi_2(g) = w_3/20 - 4$. Hence, $w_3 = 40r$. Further, the number $\chi_1(g) = (2w_1 - 10r - 25)/13$ is odd, therefore $w_1 = 13m + 6 + 5r$. Finally, $\alpha_2(g) = 0$ (if $d(u, u^g) = 2$, then the only vertex from $[u] \cap [u^g]$ belongs to Ω, a contradiction). Therefore $\alpha_1(g) = 10r + 26m + 12 = 1600 - 80r$.

Let $p = 5$. Then $w_1 + w_2 + w_3 = 320$ and $\chi_2(g) = w_3/8 - 4$. Hence, $w_3 = 40l$. Finally, $\chi_1(g) = (5w_1 - 25l - 25)/13$, therefore $w_1 = 13n + 5l + 5$. Statement (1) is proved.

Let Ω be an n-clique. If $n = 1$, then p divides 39 and 315, therefore $p = 3$. We have $\chi_1(g) = (8\alpha_1(g) - \alpha_3(g) - 156)/104$, $\chi_2(g) = (4 + \alpha_3(g))/40 - 4$. Therefore the number $(4 + \alpha_3(g))/40$ is congruent to 1 modulo 3, $\alpha_3(g) = 120l + 36$ and the number $\chi_1(g) = (\alpha_1(g) - 15l - 24)/13$ is divisible by 3. Hence, $\alpha_1(g) = 15l + 24 + 39m$.

If $n > 1$, then p divides $4 - n$ and 36, therefore either $n = 2$, $p = 2$, or $n = 4$, $p = 2, 3$. In the first case the number $\chi_2(g) = (8 + \alpha_3(g))/40 - 4$ is even and $\alpha_3(g) = 80l - 8$. Further, the number $\chi_1(g) = (\alpha_1(g) - 10l)/13 - 1$ is odd and $\alpha_1(g) = 10l + 26m$. In the second case $\chi_2(g) = (16 + \alpha_3(g))/40 - 4$ and either $p = 2$, $\alpha_3(g) = 80l - 16$, or $p = 3$ and $\alpha_3(g) = 120l + 24$. Further, $\chi_1(g) = (176 + 8\alpha_1(g) - \alpha_3(g))/104 - 25/13$ and either $p = 2$, $\alpha_1(g) = 10l + 26m + 14$, or $p = 3$ and $\alpha_1(g) = 10l + 39m + 1$, l is congruent to -1 modulo 3.

Let Ω consists of n vertices at distance 3. As $p^3_{13} = 3$, $p^3_{33} = 44$, then p divides 3 and $46 - n$. Hence, $p = 3$ and $n \in \{4, 7, 10, ..., 40\}$. We have $\chi_2(g) = (4n + \alpha_3(g))/40 - 4$ and the number $(4n + \alpha_3(g))/40$ is congruent to 1 modulo 3, therefore $\alpha_3(g) = 120l + 40 - 4n$. Further, the number $\chi_1(g) = (6n + \alpha_1(g) - 15l - 30)/13$ is divisible by 3 and $\alpha_1(g) = 15l + 30 + 39m - 6n$.

Let Ω contains an edge and does not contain vertices at distance 2 in Γ. Then Ω is an union of isolated cliques, any two vertices from distinct cliques are at distance 3 in Γ. As orders of these cliques are at most 4, then $p \leq 3$. If $p = 3$, then the orders of these cliques are equal to 1 or 4. If $p = 2$, then the orders of these cliques are equal to 2 or 4. $\square$

Lemma 9. *If* Ω *contains vertices* a, b *at distance* 2 *in* Γ*, then* $p \leq 3$.

P r o o f. Let Ω contains vertices a, b at distance 2 in Γ and Ω_0 is a connected component of Ω containing a, b.

Assume that the diameter of graph Ω_0 is equal to 2. Then by [1, 1.17.1] one of the following statements holds:

(i) $\Omega_0 \subseteq a^\perp$ and $\Omega_0(a)$ is an union of isolated cliques;

(ii) Ω_0 ia a strongly regular graph;

(iii) Ω_0 is a biregular graph with degrees of vertices α, β, where $\alpha < \beta$, and if A and B are sets of vertices from Ω_0 with degrees α and β, then A is a coclique, the lines between A and B have order 2, the lines from B have order $l = \beta - \alpha + 2 > 2$, and $|\Omega_0| = \alpha\beta + 1$.

Last case is impossible because $c_2 = 1$ in Γ.

In the case (i) we have $p \in \{2, 3\}$ because of $p^1_{33} = 12$.

In the case (ii) either $p = 2$ and Ω_0 is the pentagon, Petersen graph or Hoffman-Singletone graph, or $p > 2$ and Ω_0 is a strongly regular graph with parameters $(v', k', 2, 1)$.

Let $p > 2$. Then $\Omega(a)$ consists of e isolated triangles and either $e = 1$, $p = 3$, or $e = 2$, $p = 3, 11$, or $e = 3$, $p = 3, 5$, or $e = 4$, $p = 3$, or $e = 5$, $p = 3$, or $e = 6$, $p = 3, 7$, or $e = 7$, $p = 3$, or $e = 8$, $p = 3, 5$, or $e \geq 9$, $p = 3$.

In case $p = 11$ graph Ω is a regular graph of degree 6, $|\Omega \cap \Gamma_2(a)| = 18$, $|\Omega \cap \Gamma_3(a)| = 24$ and $|\Gamma_3(a) - \Omega|$ is not divisible by 11.

In case $p = 7$ graph Ω is a regular graph of degree 18, $|\Omega \cap \Gamma_2(a)| = 270$, $|\Omega \cap \Gamma_3(a)| = 270 \cdot 4/15 = 64$ and $|\Gamma_3(a) - \Omega|$ is not divisible by 7.

In case $p = 5$ graph Ω contains vertices of degrees 9 and 24. Assume that $|\Omega(a)| = 24$, $\Omega(a)$ contains β vertices of degree 24 in Ω and $\Omega_3(a)$ contains γ vertices of degree 24 in Ω. Then the number $21\beta + 6(24 - \beta) = |\Omega \cap \Gamma_2(a)|$ is congruent to 4 modulo 5 and $4|\Omega \cap \Gamma_2(a)| = 21\gamma + 6(|\Omega \cap \Gamma_3(a)| - \gamma)$. Hence, $|\Gamma_2(a) \cap \Omega| = (144 + 15\beta)$ and $576 + 60\beta = 15\gamma + 6|\Omega \cap \Gamma_3(a)|$, a contradiction with the fact that $|\Omega \cap \Gamma_3(a)|$ is divisible by 5.

So, Ω is an amply regular graph with parameters $(v', 9, 2, 1)$, $54 = |\Omega \cap \Gamma_2(a)|$ and $|\Omega \cap \Gamma_3(a)| = 36$. Again we have a contradiction with the fact that $|\Omega \cap \Gamma_3(a)|$ is divisible by 5.

The lemma is proved. □

Theorem 1 follows from Lemmas 8–9.

3. Proof of Corollary 1

Until the end of the paper we will assume that Γ is a distance-regular graph with intersection array $\{39, 36, 4; 1, 1, 36\}$ and the nonsolvable group $G = \mathrm{Aut}(\Gamma)$ acts transitively on the set of vertices of this graph. For the vertex $a \in \Gamma$ we get $|G : G_a| = 1600$. In view of Theorem 1 we have $p \in \{2, 3, 5\}$. Let $\bar{T}$ be the socle of the group $\bar{G} = G/O_{5'}(G)$.

Lemma 10. *If* f *is an element of order 5 of* G*,* g *is an element of order* $p < 5$ *of* $C_G(f)$ *and* $\Omega = \mathrm{Fix}(g)$*, then one of the following statements holds*:

(1) Ω *is an empty graph,* $p = 2$*,* $\alpha_3(g) = 80r$*,* $r \leq 19$*,* $\alpha_1(g) = 10r + 26m + 12 = 1600 - 80r$*, and* $m \in \{-7, -2, 3, 8, ..., 58\}$;

(2) Ω *consists of* n *vertices at distance* 3 *in* Γ*,* $p = 3$*,* $n \in \{10, 25, 40\}$*,* $\alpha_3(g) = 120l + 40 - 4n$*,* $\alpha_1(g) + \alpha_3(g) = 135l - 10n + 39m + 70 \leq 1600$ *and* m *is divisible by* 5;

(3) $p = 3$*,* $\alpha_3(g) = 120s$*,* $\alpha_0(g) = 30t + 10$*,* $\alpha_1(g) = 39l - 165t + 15s - 30$ *or* $\alpha_3(g) = 120s + 60$*,* $\alpha_0(g) = 30t - 5$*,* $\alpha_1(g) = 195l - 165t + 15s + 60$;

(4) $p = 2$*,* $\alpha_3(g) = 80s - 4\alpha_0(g)$ *and* $\alpha_1(g) = 10s + 26l + 38 - 6\alpha_0(g)$*,* l *is congruent to* 2 *modulo* 5.

P r o o f. In view of Theorem 1 Fix(f) is empty graph, $\alpha_1(f) = 65n+10l+10$ and $\alpha_3(f) = 200l$.

If Ω is an empty graph, then $p = 2$, $\alpha_3(g) = 80r$ and $\alpha_1(g) = 10r + 26m + 12 = 1600 - 80r$ is divisible by 5. Hence, $13m + 6$ is divisible by 5 and $m \in \{-7, -2, 3, 8, ..., 58\}$. Finally, $26m + 12 = 1600 - 90r$, therefore m is congruent to 2 modulo 3 and $m \in \{-7, 8, 23, 38, 53\}$.

If Ω is an n-clique, then n is divisible by 5, we have got a contradiction.

If Ω consists of n vertices at distance 3 in Γ, then $p = 3$, $n \in \{10, 25, 40\}$, the numbers $\alpha_3(g) = 120l + 40 - 4n$ and $\alpha_1(g) = 15l + 30 + 39m - 6n$ are divisible by 5. Hence, m is divisible by 5, $\alpha_1(g) + \alpha_3(g) = 135l - 10n + 39m + 70 \leq 1600$.

If $p = 3$, then $\chi_2(g) = (4\alpha_0(g)+\alpha_3(g))/40-4$ and the number $(4\alpha_0(g)+\alpha_3(g))/40$ is congruent to 1 modulo 3. Further, the number $\chi_1(g) = (44\alpha_0(g) + 8\alpha_1(g) - \alpha_3(g))/104 - 25/13$ is divisible by 3, $\alpha_3(g)$ is divisible by 60. If $\alpha_3(g) = 120s$, then $\alpha_0(g) = 30t+10$, $\alpha_1(g) = 39l-165t+15s-30$. If $\alpha_3(g) = 120s + 60$, then $\alpha_0(g) = 30t - 5$, $\alpha_1(g) = 195l - 165t + 15s + 60$.

If $p = 2$, then $\chi_2(g) = (4\alpha_0(g) + \alpha_3(g))/40 - 4$, $4\alpha_0(g) + \alpha_3(g) = 80s$. Further, $\alpha_1(g) = -6\alpha_0(g) + 10s + 26l + 38$ and $13l + 19$ is divisible by 5, therefore $l \in \{2, 7, ...\}$. Finally, $1600 - 5\alpha_0(g) + 80s = -6\alpha_0(g) + 10s + 26l + 38$, $1600 = -70s - \alpha_0(g) + 26l + 38$. □

Lemma 11. *The following statements hold*:

(1) $\bar{T} = L \times M$, *and each of subgroups* L, M *is isomorphic to one of the following groups* Z_5, A_5, A_6 *or* $PSp(4,3)$;

(2) *in case* $|\bar{T} : \bar{T}_a| = 40^2$ *we have* $O_{5'}(G) = 1$ *and this case is realized if one of the following statements holds*:

(*i*) $L \cong M \cong PSp(4,3)$, *or*

(*ii*) $L \cong PSp(4,3)$, $|L : L_a| = 40$, $M \cong A_6$ *and* $|M_a| = 9$, *or*

(*iii*) $L \cong M \cong A_6$ *and* $|L_a| = |M_a| = 9$.

P r o o f. Recall that a nonabelian simple $\{2,3,5\}$-group is isomorphic to A_5, A_6 or $PSp(4,3)$ (see, [6, Table 1]). Hence, in view of Theorem 1 we have $\bar{T} = L \times M$, each of subgroups L, M is isomorphic to one of the following groups A_5, A_6 or $PSp(4,3)$.

If $\bar{T} \cong PSp(4,3)$, then the group $\bar{T}_a$ has an index 40 in $\bar{T}$ and is isomorphic to $E_9.SL_2(3)$ or $E_{27}.S_4$.

If $\bar{T} \cong A_6$, then the group $\bar{T}_a$ has an index in $\bar{T}$, divisible by 10, and dividing 40.

If $\bar{T} \cong A_5$, then the group $\bar{T}_a$ has an index in $\bar{T}$, divisible by 10, and dividing 20.

In case $|\bar{T} : \bar{T}_a| = 40^2$ we have $O_{5'}(G) = 1$ and this case is realized if one of the following statements holds: either $L \cong M \cong PSp(4,3)$, or $L \cong PSp(4,3)$, $M \cong A_6$ $|M_a| = 9$, or $L \cong M \cong A_6$ and $|L_a| = |M_a| = 9$. □

Corollary is proved.

4. Conclusion

We found possible automorphisms of a distance-regular graph with intersection array $\{39, 36, 4; 1, 1, 36\}$. In particular this graph is not arc-transitive.

REFERENCES

1. Brouwer A. E., Cohen A. M., Neumaier A. *Distance-Regular Graphs.* New York: Springer-Verlag, 1989. 495 p. DOI: 10.1007/978-3-642-74341-2
2. Makhnev A. A., Nirova M. S. On distance-regular graphs with $\lambda = 2$. *J. Sib. Fed. Univ. Math. Phys.*, 2014. Vol. 7, No. 2. P. 204–210.

3. Behbahani M., Lam C. Strongly regular graphs with nontrivial automorphisms. *Discrete Math.*, 2011. Vol. 311, No. 2–3. P. 132–144. DOI: 10.1016/j.disc.2010.10.005
4. Cameron P. J. *Permutation Groups.* London Math. Soc. Student Texts, No. 45. Cambridge: Cambridge Univ. Press, 1999.
5. Gavrilyuk A. L., Makhnev A. A. On automorphisms of distance-regular graph with the intersection array $\{56, 45, 1; 1, 9, 56\}$. *Doklady Mathematics*, 2010. Vol. 81, No. 3. P. 439–442. DOI: 10.1134/S1064562410030282
6. Zavarnitsine A. V. Finite simple groups with narrow prime spectrum. *Sib. Electron. Math. Izv.*, 2009. Vol. 6. P. S1–S12.

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 79–87

DOI: 10.15826/umj.2018.2.009

ALTRUISTIC AND AGGRESSIVE TYPES OF BEHAVIOR IN A NON-ANTAGONISTIC DIFFERENTIAL GAME

Anatolii F. Kleimenov

N.N. Krasovskii Institute of Mathematics and Mechanics
Ural Branch of Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, Russia, 620990;

Ural Federal University, 19 Mira str., Ekaterinburg, Russia, 620002

kleimenov@imm.uran.ru

Abstract: An example of a non-antagonistic positional (feedback) differential two-person game (NPDG) is considered in which each of two players, in addition to the normal type of behavior, oriented toward maximizing own functional, can use other types of behavior. In particular, it can be altruistic and aggressive types. In the course of the game players can switch their behavior from one type to other. The use by players of types of behavior other than normal can lead to outcomes more preferable for them than in a game with only normal behavior. The example with the dynamics of simple motion on a plane and phase constraints illustrates the procedure of constructing new solutions.

Keywords: Non-antagonistic positional differential game, Altruistic type of behavior, Agressive type of behavior.

Introduction

In this article we consider a non-antagonistic positional differential two-person game (see, for example, [9]), for which emphasis is placed on the case where each of the two players, in addition to the normal (*nor*), type of behavior, oriented on maximizing their own functional, can use other types of behavior introduced in [2, 5], such as altruistic (*alt*), aggressive (*agg*) types. It is assumed that during the game, players can switch their behavior from one type to another. The idea of using the players to switch their behavior from one type to another in the course of the game was applied to the game with cooperative dynamics in [5] and for the repeated bimatrix 2×2 game in [3], which allowed to obtain new solutions in these games.

It is assumed that in the game each player chooses the indicator function determined over the whole time interval of the game and takes values in the set $\{nor, alt, agg\}$, along with the choice of the positional strategy. Player's indicator function shows the dynamics for changing the type of behavior that this player adheres to. Rules for the formation of controls are introduced for each pair of behaviors of players.

The formalization of positional strategies in the game is based on the formalization and results of the general theory of antagonistic positional differential games [7, 8]. For non-antagonistic positional differential games this formalization was developed in [1].

In the article the concept of the BT-solution is introduced.

An example of a game with dynamics of simple motion in the plane and phase constraints is proposed. We assume that the first and second players can exhibit altruism and aggression towards their partner for some time periods and a case of mutual aggression is allowed. Sets of BT- solutions are described. This paper is a continuation of [4].

1. Equations of motion and phase constraints.

Let equations of dynamics be as follows

$$\dot{x} = u + v, \quad x, u, v \in R^2, \quad \|u\| \leq 1, \quad \|v\| \leq 1, \quad 0 \leq t \leq \vartheta, \quad x(0) = x^0, \tag{1.1}$$

where x is the phase vector; u and v are controls of Player 1 (P1) and Player 2 (P2), respectively. Let payoff functional of Player i be

$$I_i = \sigma_i(x(\vartheta)) = M - \|x(\vartheta) - a^{(i)}\|, \quad i = 1, 2, \tag{1.2}$$

where M is a constant. That is, the goal of Player i is to bring vector $x(\vartheta)$ as close as possible to the target point $a^{(i)}$.

Let the initial conditions and values of parameters be given (Fig. 1):

$$\vartheta = 5.0, \quad x^0 = (0, 0), \quad a^{(1)} = (10, 8), \quad a^{(2)} = (-10, 8), \quad M = 18.$$

The game has the following phase restrictions. The trajectories of the system (1.1) are forbidden from entering the interior of the set S, which is obtained by removing from the quadrilateral $Oabc$ the line segment Oe (Fig. 1). The set S consists of two parts S_1 and S_2, that is, $S = S_1 \cup S_2$. Coordinates of the points defining the phase constraints: $a = (-4.5, 3.6)$, $b = (0, 8)$, $c = (6.5,\ 5.2)$, $O = (0, 0)$, $e = (3.25,\ 6.6)$. We have $a \in Oa^{(2)}$, $c \in Oa^{(1)}$, $e \in bc$ and $|a^{(1)}b| = |a^{(2)}b| = 10$.

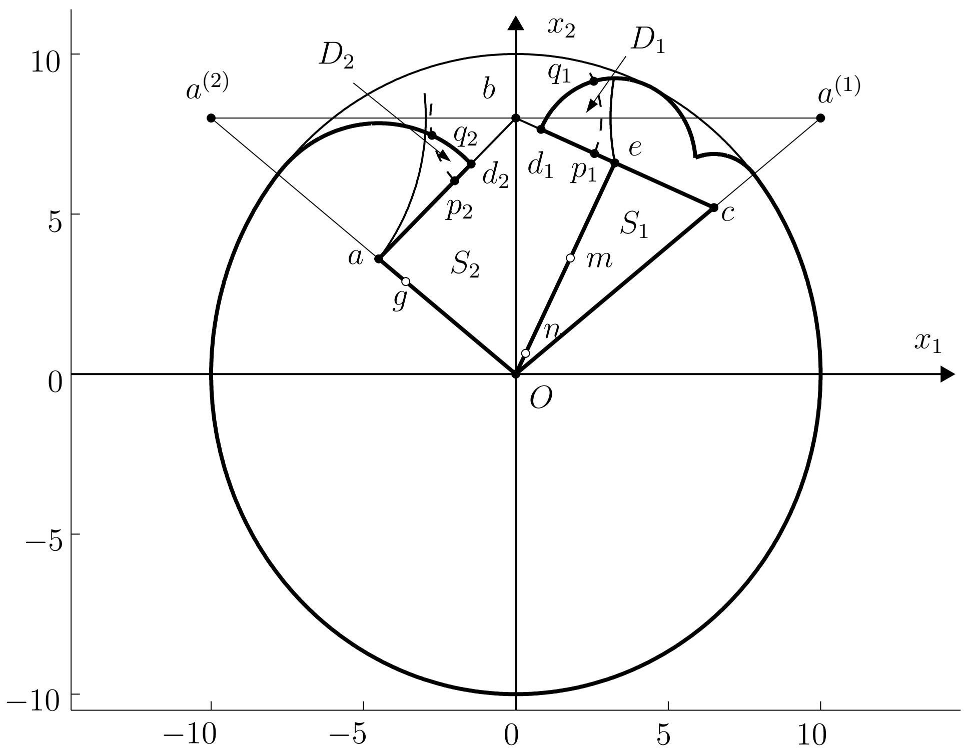


Figure 1. The attainability set

2. Attainability set

Attainability set of the system (1.1) constructed for the moment ϑ consists of points of the circle of radius 10 located not higher than the three-link segment aOc and also bounded by two arcs connecting the large circle with the sides ab and bc of the quadrilateral. The first arc is an arc of the circle with center at the point a and radius $r_1 = 10 - |Oa| = |ad_2|$. The second (composite) arc consists of an arc of the circle with center at the point e and radius $r_2 = 10 - |Oe| = |ed_1|$ and an arc of the circle with center at the point c and radius $r_3 = 10 - |Oc|$. (Fig. 1). Results of approximate calculations: $r_1 = 4.24$, $r_2 = 2.64$, $r_3 = 1.68$, $d_1 = (0.82, 7.65)$, $d_2 = (-1.47, 6.56)$. We have also: $|Oa^{(1)}| = |Oa^{(2)}| = 12.81$.

In Fig. 1 the dashed lines represent arcs of the circle with center at the point b and radius $r_4 = |Oa^{(1)}| - |a^{(1)}b| = 12.81 - 10 = 2.81$. These arcs intersect the sides ab and bc at the points $p_1 = (2.58, 6.89)$ and $p_2 = (-2.01, 6.04)$. By construction, the lengths of the two-links $a^{(1)}bp_2$ and $a^{(2)}bp_1$ are equal to each other and equal to the lengths of the segments $Oa^{(1)}$ and $Oa^{(2)}$.

3. Strategies and motions

Assume that both players have information about the current position of the game $(t, x(t))$. Then P1 and P2 acts in the class of pure positional strategies.

The formalization of players' positional strategies and motions generated by them in a NPDG is based on the formalization and results of antagonistic positional differential games theory from the books [7, 8]. The following presentation is given in [1].

Strategy of P1 is identified with the pair $U = \{u(t, x, \varepsilon),\ \beta_1(\varepsilon)\}$, where $u(\cdot)$ is an arbitrary function of position (t, x) and a positive precision parameter $\varepsilon > 0$ and taking values in the set $\{u \in R^2, \|u\| \le 1\}$. The function $\beta_1 : (0, \infty) \longmapsto (0, \infty)$ is a continuous monotonic function satisfying the condition $\beta_1(\varepsilon) \to 0$ if $\varepsilon \to 0$. For a fixed ε the value $\beta_1(\varepsilon)$ is the upper bound step of subdivision the segment $[t_0, \vartheta]$, which P1 applies when forming step-by-step motions.
Similarly, the strategy of P2 is defined as $V = \{v(t, x, \varepsilon), \beta_2(\varepsilon)\}$.

Motions of two types — approximated (step-by-step) and ideal (limiting) are considered as motions generated by a pair {strategy of P1 – strategy of P2}.

Approximated motion $x^{\varepsilon}_{\Delta}[\cdot] = x[\cdot, t_0, x_0, U, \varepsilon_1, \Delta_1, V, \varepsilon_2, \Delta_2]$ generated by a pair of strategies (U, V) from the initial position (t_0, x_0) for fixed values of the players' precision parameters ε_1 and ε_2, for fixed subdivisions $\Delta_1 = \{t_i^{(1)}\}$ and $\Delta_2 = \{t_i^{(2)}\}$ of the interval $[t_0, \vartheta]$ chosen by P1 and P2, respectively, under the conditions $\delta(\Delta_i) \le \beta_i(\varepsilon_i), i = 1, 2$, is introduced as step-by-step solution of the differential equation

$$\begin{aligned}
&\dot{x}^{\varepsilon}_{\Delta}[t] = f(t, x^{\varepsilon}_{\Delta}[t],\, u^{\varepsilon_1}_{\Delta_1}[t], v^{\varepsilon_2}_{\Delta_2}[t]), \quad x^{\varepsilon}_{\Delta}[t_0] = x_0,\\
&u^{\varepsilon_1}_{\Delta_1}[t] = u(t_i^{(1)}, x^{\varepsilon}_{\Delta}[t_i^{(1)}], \varepsilon_1), \quad t_i^{(1)} \le t < t_{i+1}^{(1)},\\
&v^{\varepsilon_2}_{\Delta_2}[t] = v(t_j^{(2)}, x^{\varepsilon}_{\Delta}[t_j^{(2)}], \varepsilon_2), \quad t_j^{(2)} \le t < t_{j+1}^{(2)}.
\end{aligned}$$

A limiting motion generated by the pair of strategies (U, V) from the initial position (t_0, x_0) is a continuous function $x[t] = x[t, t_0, x_0, U, V]$ for which there exists a sequence of approximated motions $\{x[t, t_0^k, x_0^k, U, \varepsilon_1^k, \Delta_1^k, V, \varepsilon_2^k, \Delta_2^k]\}$, uniformly converging to $x(t)$ on $[t_0, \vartheta]$ as $k \to \infty$, $\varepsilon_1^k \to 0$, $\varepsilon_2^k \to 0$, $t_0^k \to t_o$, $x_0^k \to x_o$.
A pair of strategies (U, V) generates a nonempty compact (in the metric of the space $C[t_0, \vartheta]$) set $X(t_0, x_0, U, V)$ consisting of limit motions $x[\cdot, t_0, x_0, U, V]$.

The control laws $(U, \varepsilon_1, \Delta_1)$ and $(V, \varepsilon_2, \Delta_2)$ are said to be consistent with respect to the precision parameter if $\varepsilon_1 = \varepsilon_2$. Consistent control laws generate consistent approximate motions, the sequences of which generate consistent limit motions.

4. Auxiliary zero-sum positional differential games Γ_i

Consider auxiliary zero-sum positional differential games Γ_1 and Γ_2. Dynamics of both games is described by the equation (1.1). In the game Γ_i Player i maximizes the payoff functional $\sigma_i(x(\vartheta))$ (1.2) and Player $3-i$ opposes him. It follows from [7, 8] that both games Γ_1 and Γ_2 have universal saddle points

$$\{u^{(i)}(t,x,\varepsilon),\, v^{(i)}(t,x,\varepsilon)\}, \quad i=1,2, \tag{4.1}$$

and continuous value functions

$$\gamma_1(t,x), \quad \gamma_2(t,x). \tag{4.2}$$

The property of strategies (4.1) to be universal means that they are optimal not only for the fixed initial position (t_0, x_0) but also for any position (t_*, x_*) assumed as initial one.

It is not difficult to see that the value of $\gamma_i(t,x)$ (4.2) is the maximal guaranteed payoff of the Player i in the position (t,x) of the game.

The value functions $\gamma_1(t,x)$ and $\gamma_2(t,x)$, $0 \le t \le \vartheta$, $x \in R^2 \backslash S$ in our example will be as follows

$$\gamma_i\left(t,x\right) = \begin{cases} 18 - \|(x - a^{(i)}\|, & xa^{(i)} \bigcap intS = \emptyset, \\ 18 - \rho_S\left(x, a^{(i)}\right) & otherwise, \end{cases} \tag{4.3}$$

where $i=1,2$, and $\rho_S\left(x, a^{(i)}\right)$ denotes the smallest of the two distances from the point x to the point $a^{(i)}$, one of which is calculated when the set S is bypassed clockwise and the other when the set S is bypassed counterclockwise.

5. NE- and $P(NE)$-solutions in NPDG

At first we solve the game NPDG (without abnormal behavior types). Introduce the following definitions from [1].

Definition 1. *A pair of strategies (U^N, V^N) is called a Nash equilibrium solution (NE-solution) of the game, if for any motion $x^*[\cdot] \in X(t_0, x_0, U^N, V^N)$, any moment $\tau \in [t_0, \vartheta]$, and any strategies U and V the following inequalities hold:*

$$\max_{x[\cdot]} \sigma_1(x[\vartheta, \tau, x^*[\tau], U, V^N]) \le \min_{x[\cdot]} \sigma_1(x[\vartheta, \tau, x^*[\tau], U^N, V^N]),$$
$$\max_{x[\cdot]} \sigma_2(x[\vartheta, \tau, x^*[\tau], U^N, V]) \le \min_{x[\cdot]} \sigma_2(x[\vartheta, \tau, x^*[\tau], U^N, V^N]).$$

where the operations min are performed over a set of agreed motions, and the operations max by sets of all motions.

Definition 2. *An NE-solution (U^P, V^P) which is Pareto non-improvable with respect to the values I_1, I_2 (1.2) is called a $P(NE)$-solution.*

In [1] the following structure of NE- and $P(NE)$-solutions was established. Namely, it was shown that all NE- and $P(NE)$-solutions of the game NPDG can be found in the class of pairs of strategies (U, V) each of which generates a unique limit motion (trajectory). The decision strategies that make up such a pair generating the trajectory $x^*(\cdot)$ have the form

$$U^0 = \{u^0(t,x,\varepsilon), \beta_1^0(\varepsilon)\}, \quad V^0 = \{v^0(t,x,\varepsilon), \beta_2^0(\varepsilon)\}, \tag{5.1}$$

$$u^0(t,x,\varepsilon) = \begin{cases} u^*(t,\varepsilon), & \|x - x^*(t)\| < \varepsilon\varphi(t), \\ u^{(2)}(t,x,\varepsilon), & \|x - x^*(t)\| \geq \varepsilon\varphi(t), \end{cases}$$

$$v^0(t,x,\varepsilon) = \begin{cases} v^*(t,\varepsilon), & \|x - x^*(t)\| < \varepsilon\varphi(t), \\ v^{(1)}(t,x,\varepsilon), & \|x - x^*(t)\| \geq \varepsilon\varphi(t), \end{cases}$$

for all $t \in [t_0, \vartheta]$, $\varepsilon > 0$. In (5.1) we denote by $u^*(t,\varepsilon)$, $v^*(t,\varepsilon)$ families of program controls generating the limit motion $x^*(t)$. The function $\varphi(\cdot)$ and the functions $\beta_1^0(\cdot)$ and $\beta_2^0(\cdot)$ are chosen in such a way that the approximated motions generated by the pair (U^0, V^0) from the initial position (t_0, x_0) do not go beyond the $\varepsilon\varphi(t)$-neighborhood of the trajectory $x^*(t)$. Functions $u^{(2)}(\cdot,\cdot,\cdot)$ and $v^{(1)}(\cdot,\cdot,\cdot)$ are defined in (4.1).

It is proved in [1] that the point $t = \vartheta$ is the maximum point of the value function $\gamma_i(t,x)$, $i = 1, 2$ computed along NE-trajectory and $P(NE)$-trajectory.

One can check that in this game the trajectory $x(t) \equiv 0$, $t \in [0, 5]$ (stationary point O) is the only NE-trajectory, and, consequently, the only $P(NE)$-trajectory; the players' gains on it are $I_1 = I_2 = 5.19$.

6. Types of behavior

Let us move on to the game NPDGwBT (with abnormal behavior types), in which each player during certain periods of time may exhibit altruism and aggression towards another player, and the mutual aggression is allowed. The definitions of behavior types other than normal are given in [2, 5].

Definition 3. *We say that on the time interval $(\tau_1, \tau_2) \subset [t_0, \vartheta]$ Player 1 adheres to the altruistic (alt) type of behavior with respect to Player 2, if his payoff functional on this interval is equal to the functional I_2 (1.2) of Player 2.*

Definition 4. *We say that on the time interval $(\tau_1, \tau_2) \subset [t_0, \vartheta]$ Player 1 adheres to the aggressive (agg) type of behavior with respect to Player 2, if his payoff functional on this interval is equal to the functional $-I_2$, where I_2 (1.2) is the payoff functional of Player 2.*

Definition 5. *We will say that on the time interval $(\tau_1, \tau_2) \subset [t_0, \vartheta]$ Player 1 adheres to the paradoxical (par) type of behavior if his payoff functional on this interval is equal to the functional I_1 (1.2), taken with the opposite sign.*

Similarly, we define the altruistic and aggressive types of behavior for Player 2 towards Player 1, as well as the paradoxical type of behavior for Player 2.

In this paper, we do not use the paradoxical type of players behavior. Note that the aggressive type of player behavior is actually used in NPDG in the form of punishment strategies contained in the structure of the game's decisions (see, for example, [1, 6]).

The above definitions characterize the extreme types of behavior of players. In reality, however, real individuals behave, as a rule, partly normal, partly altruistic, partly aggressive and partly paradoxical. In other words, mixed types of behavior seem to be more consistent with reality.

If each player is confined to "pure" types of behavior, then in the game (1.1), (1.2) there are 9 possible pairs of types of behavior: (nor, nor), (nor, alt), (nor, agg), (alt, nor), (alt, alt), (alt, agg), (agg, nor), (agg, alt), (agg, agg). For two pairs (nor, alt) and (alt, nor) the interests of the players coincide and they solve a team problem of control. For two pairs (nor, agg) and (agg, nor) players have opposite interests and, therefore, they play a zero-sum game. The remaining 5 pairs define a non-antagonistic games.

The idea of using the players to switch their behavior from one type to another in the course of the game was applied to the game with cooperative dynamics in [5] and for the repeated bimatrix 2×2 game in [3], which allowed to obtain new solutions in these games.

The extension of this approach to non-antagonistic positional differential games leads to new formulation of problems. In particular, it is of interest to see how the player's winnings, obtained on Nash solutions, are transformed. The actual task is to minimize the time of "abnormal" behavior, provided that the players' payoffs are greater than when the players behave normally.

7. Formalization of actions. Rules 1,2

In NPDGwBT we assume that simultaneously with the choice of positional strategy [7, 8] each player also chooses his indicator function [5]. We denote the indicator function of Player i by the symbol $\alpha_i : [t_0, \vartheta] \longmapsto \{nor, alt, agg\}$, $i = 1, 2$. If the indicator function of some player takes a value, say, alt, on some time interval, then this player acts on this interval as an altruist in relation to his partner. Thus, in the game NPDGwBT P1 controls the choice of a pair of *actions* {position strategy, indicator function} $(U,\ \alpha_1(\cdot))$, and P2 controls the choice of a pair of actions $(V,\ \alpha_2(\cdot))$.

As mentioned above, for any pair of types of behavior three types of decision making problems can arise: a team problem, a zero-sum game, and a non-antagonistic game. We adopt the following Rules 1–2.

Rule 1. If on the time interval $(\tau_1, \tau_2) \subset [t_0, \vartheta]$ the players' indicator functions generate a non-antagonistic game, then on this interval P1 and P2 choose one of $P(NE)$-solutions of this game. If a zero-sum game is realized, then as a solution, P1 and P2 choose one of saddle points of this game. Finally, if a team problem of control is realized, then P1 and P2 choose one of the pairs of controls such that the value function $\gamma_i(t, x)$ calculated along the generated trajectory is non-decreasing function, where i is the number of the player whose functional is maximized in team problem.

Generally speaking, the same part of the trajectory can be tracked by several pairs of players' types of behavior, and these pairs may differ from each other by the time of use of abnormal types. It is natural to introduce the following Rule 2.

Rule 2. If there are several pairs of types of behavior that track a certain part of the trajectory, then P1 and P2 choose one of them that minimizes the time of using abnormal types of behavior.

8. BT-solution in NPDGwBT

We now introduce the definition of the solution of the game NPDGwBT. The definition of BT-solution is given in [5].

Note that the set of motions generated by a pair of actions $\{(U, \alpha_1(\cdot)), (V, \alpha_2(\cdot))\}$ coincides with the set of motions generated by the pair (U, V) in the corresponding NPDG.

Definition 6. *The pair $\{(U^0, \alpha_1^0(\cdot)), (V^0, \alpha_2^0(\cdot))\}$, consistent with Rules 1, 2, forms a BT-solution of the game NPDGwBT if there exists a trajectory $x^{BT}(\cdot)$ generated by this pair and there* *is a $P(NE)$-solution in the corresponding game NPDG generating the trajectory $x^P(\cdot)$ such that the following inequalities are true*

$$\sigma_i(x^{BT}(\vartheta)) \geq \sigma_i(x^P(\vartheta)), \quad i = 1, 2,$$

where at least one of the inequalities is strict.

Definition 7. *The BT-solution $\{(U^0, \alpha_1^0(\cdot)), (V^0, \alpha_2^0(\cdot))\}$, which is Pareto non-improvable with respect to the values I_1, I_2 (1.2), is called $P(BT)$-solution of the game NPDGwBT.*

Problem 1. Find the set of BT-solutions.

Problem 2. Find the set of $P(BT)$-solutions.

In the general case, Problems 1 and 2 have no solutions. However, it is quite expected that the use of abnormal behavior types by players in the game NPDGwBT can in some cases lead to outcomes more preferable for them than in the corresponding game NPDG only with the normal type of behavior.

In our example, just such a situation will take place.

9. Building BT-solutions

Let us now turn to the game NPDGwBT, in which each player during certain periods of time may exhibit altruism and aggression towards another player, and the case of mutual aggression is allowed.

In the attainability set, we find all the points x for which the inequalities hold

$$\begin{gathered} \sigma_i(x) \geq \sigma_i(O), \quad i = 1, 2, \\ \sigma_1(x) + \sigma_2(x) > \sigma_1(O) + \sigma_2(O). \end{gathered} \tag{9.1}$$

Such points form two sets D_1 and D_2 (see Fig. 1). The set D_1 is bounded by the segment p_1d_1, and also by the arcs p_1q_1 and q_1d_1 of the circles mentioned above. The set D_2 is bounded by the segment p_2d_2, and also by the arcs d_2q_2 and q_2p_2 of the circles mentioned. On the arc p_1q_1, the non-strict inequality (9.1) for $i = 2$ becomes an equality, and on the arc q_2p_2, the non-strict inequality (9.1) becomes an equality for $i = 1$. At the remaining points sets D_1 and D_2 , the non-strict inequalities (9.1) for $i \in \{1, 2\}$ are strict.

We construct a BT-solution, leading to the point $d_1 \in D_1$. Let us find a point m equidistant from the point $a^{(2)}$ if we go around the set S_2 clockwise, or if we go around S_2 counterclockwise. We also find a point n equidistant from the point $a^{(1)}$ as if we were go around the set S_1 clockwise, or if we go around S_1 counterclockwise. The results of the calculations: $m = (1.79, 3.63)$, $n = (0.32, 0.65)$.

Consider the trajectory Oed_1; the players' gains on it are $I_1 = 8.82$, $I_2 = 7.10$, that is, the gains of both players on this trajectory are greater than the gains on the single $P(NE)$ -trajectory. As follows from the above, the trajectory Oed_1 is not Nash one. Therefore, if it is possible to construct indicator functions-programs of players that provide motion along this trajectory, then a BT-solution will be constructed.

First of all find that if we move along the trajectory Oed_1 with the maximum velocity for $t \in [0, 5]$, the time to hit the point n will be $t = 0.361$, the point m will be $t = 2.022$, and the point e will be $t = 3.678$. It is easy to verify that for such a motion along the trajectory Oed_1 on the interval $t \in [0, 0.361]$, both functions $\gamma_1(t, x)$ and $\gamma_2(t, x)$ (4.3) decrease monotonically;

for motion on the interval $t \in [0.361,\ 2.022]$, the function $\gamma_2(t,x)$ continues to decrease, and the function $\gamma_1(t,x)$ increases; for motion on the interval $t \in [2.022,\ 3.678]$, both functions increase; finally, on the remaining interval $t \in [3.678,\ 5]$, the function $\gamma_2(t,x)$ continues to increase, and the function $\gamma_1(t,x)$ decreases.

We check that on the segment On of the trajectory, the pair (agg, agg), which determines the non-antagonistic game, is the only pair of types of behaviors that realizes motion on the segment in accordance with Rule 1; this is the motion generated by the $P(NE)$-solution, the best for both players. In the next segment nm, two pairs of types of behaviors realize motion on the segment according to Rule 1, namely (nor, alt) and (agg, alt); however, according to Rule 2, only the pair (nor, alt) remains; it defines a team problem of control in which the motion represents the maximum shift in the direction of point m. There are already four pairs of "candidates" (nor, nor), (alt, nor), (nor, alt) and (alt, alt) on the segment me, but according to Rule 2 the last three pairs are discarded; the remaining pair defines a non-antagonistic game and the motion on this segment is generated by the $P(NE)$-solution of the game. Finally, for the last segment ed_1, the only pair of types of behaviors is the pair (alt, nor), which defines a team problem of control; the motion represents the maximum shift in the direction of the point d_1.

Thus, we have constructed the following indicator function-programs

$$\alpha_1^{(2)}(t) = \{agg, t \in [0, 0.361);\ nor, t \in [0.361, 3.678);\ alt, t \in [3.678, 5]\}, \tag{9.2}$$

$$\alpha_2^{(2)}(t) = \{agg, t \in [0, 0.361);\ alt, t \in [0.361, 2.022);\ nor, t \in [2.022, 5]\}. \tag{9.3}$$

We denote by $(U^{(2)}, V^{(2)})$ the pair of players' strategies that generate the limit motion Oed_1 for $t \in [0, 5]$ and is consistent with the constructed indicator functions. Then we obtain the following assertion.

Theorem 1. *The pair of actions* $\{(U^{(2)}, \alpha_1^{(2)}(\cdot)), (V^{(2)}, \alpha_2^{(2)}(\cdot))\}$ (9.2), (9.3) *provides the* BT*-solution.*

Following the scheme of the proof of Theorem 1, we arrive at the following assertion.

Theorem 2. *The sets* D_1 *and* D_2 *consist of those and only those points that are endpoints of the trajectories generated by the* BT*-solutions of the game.*

10. Conclusion

In this paper, we use complex switching, namely, from one type of behavior to another, changing the nature of the problem of optimization from non-antagonistic games to zero-sum games or team problem of control and vice versa. These switchings are carried out according to pre-selected indicator function-programs. Each player controls the choice of a pair of actions {positional strategy, indicator function}. Thus, the possibilities of each player in the general case have expanded (increased) and it is possible to introduce a new concept of a game solution ($P(BT)$-solution) in which both players increase their payoffs in comparison with the payoffs in Nash equilibrium in the game without switching types of behavior. For players, it is advantageous to implement $P(BT)$-trajectory; so they will follow the declared indicator function-programs, for example, (9.2), (9.3).

REFERENCES

1. Kleimenov A. F. *Neantagonisticheskie positsionnye differentsialnye igry* [Non-antagonistic Positional Differential Games]. Ekaterinburg: Nauka, 1993. 185 p. (in Russian).
2. Kleimenov A. F. Solutions in a non-antagonistic positional differential game. *J. Appl. Math. Mech.*, 1997. Vol. 61, No. 5. P. 717–723. DOI: 10.1016/S0021-8928(97)00094-4
3. Kleimenov A. F. An approach to building dynamics for repeated bimatrix 2x2 games involving various behavior types. In: *Dynamic and Control*. London: Gordon and Breach Sci. Publ., 1998. P. 195–204.
4. Kleimenov A. F. Altruistic behavior in a non-antagonistic positional differential game. *Autom Remote Control*, 2017. Vol. 78, No. 4. P. 762–769. DOI: 10.1134/S0005117917040178
5. Kleimenov A. F., Kryazhimskii A. V. N*ormal Behavior, Altruism and Aggression in Cooperative Game Dynamics.* Interim Report IR-98-076. Laxenburg: IIASA, 1998. 47 p. URL: https://core.ac.uk/download/pdf/52947411.pdf
6. Kononenko A. F. On equilibrium positional strategies in nonantagonistic differential games. *Dokl. Akad. Nauk SSSR*, 1976. Vol. 231, No. 2. P. 285–288. (in Russian).
7. Krasovskii N. N. *Upravlenie dinamicheskoi sistemoi* [Control of a Dynamical System]. Moscow: Nauka, 1985. 520 p. (in Russian).
8. Krasovskii N. N., Subbotin A. I. *Game-Theoretical Control Problems.* New York: Springer-Verlag, 1988. 517 p.
9. Petrosyan L. A., Zenkevich N. A., Shevkoplyas E. V. *Teoriya igr* [Game Theory]. St. Petersburg: BHV-Petersburg, 2012. 424 p. (in Russian).

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 88–98

DOI: 10.15826/umj.2018.2.010

FORMATION OF VERSIONS OF SOME DYNAMIC INEQUALITIES UNIFIED ON TIME SCALE CALCULUS

Muhammad Jibril Shahab Sahir

Department of Mathematics, University of Sargodha, Sub-Campus Bhakkar, Pakistan & GHSS, 67/ML, Bhakkar, Pakistan

jibrielshahab@gmail.com

Abstract: The aim of this paper is to present some comprehensive and extended versions of classical inequalities such as Radon's Inequality, Bergström's Inequality, the weighted power mean inequality, Schlömilch's Inequality and Nesbitt's Inequality on time scale calculus. In time scale calculus, results are unified and extended. The theory of time scale calculus is applied to unify discrete and continuous analysis and to combine them in one comprehensive form. This hybrid theory is also widely applied on dynamic inequalities. The study of dynamic inequalities has received a lot of attention in the literature and has become a major field in pure and applied mathematics.

Keywords: Radon's Inequality, Bergström's Inequality, the weighted power mean inequality, Schlömilch's Inequality, Nesbitt's Inequality.

1. Introduction

The time scale calculus has a scope for many applications in the field of dynamic inequalities. The time scale calculus was initiated by Stefan Hilger as given in [11]. A *time scale* is an arbitrary nonempty closed subset of the real numbers. The time scale calculus is studied as delta calculus, nabla calculus and diamond-α calculus. Basic work on dynamic inequalities is done by Ravi Agarwal, George Anastassiou, Martin Bohner, Allan Peterson, Donal O'Regan, Samir Saker and many other authors.

We will prove the following results given in (1.1), (1.2) and (1.3) on time scales.

The inequality from (1.1) is called Bergström's Inequality as given in [4–6, 13].

Theorem 1. *If $n \in \mathbb{N}$, $x_k \in \mathbb{R}$ and $y_k > 0$, $k \in \{1, 2, ..., n\}$, then*

$$\frac{\left(\sum\limits_{k=1}^{n} x_k\right)^2}{\sum\limits_{k=1}^{n} y_k} \leq \sum_{k=1}^{n} \frac{x_k^2}{y_k}, \tag{1.1}$$

with equality if and only if

$$\frac{x_1}{y_1} = \frac{x_2}{y_2} = ... = \frac{x_n}{y_n}.$$

The upcoming result is called Radon's Inequality as given in [15].

Theorem 2. *If $n \in \mathbb{N}$, $x_k \geq 0$ and $y_k > 0$, $k \in \{1, 2, ..., n\}$ and $\beta \geq 0$, then*

$$\frac{\left(\sum\limits_{k=1}^{n} x_k\right)^{\beta+1}}{\left(\sum\limits_{k=1}^{n} y_k\right)^{\beta}} \leq \sum_{k=1}^{n} \frac{x_k^{\beta+1}}{y_k^{\beta}}. \tag{1.2}$$

Inequality (1.2) is widely studied by many authors because it has many applications.

The following inequality is generalized Radon's Inequality as given in [10].

Theorem 3. *If $n \in \mathbb{N}$, $x_k \geq 0$, $y_k > 0$, $k \in \{1, 2, ..., n\}$ and $\beta \geq \gamma \geq 0$, then*

$$n^{\gamma-\beta} \frac{\left(\sum\limits_{k=1}^{n} x_k\right)^{\beta+1}}{\left(\sum\limits_{k=1}^{n} y_k\right)^{\gamma}} \leq \sum_{k=1}^{n} \frac{x_k^{\beta+1}}{y_k^{\gamma}}, \tag{1.3}$$

with equality if and only if $x_1 = x_2 = ... = x_n$ and $y_1 = y_2 = ... = y_n$.

In this paper, it is assumed that all considerable integrals exist and are finite and $\mathbb{T}$ is a time scale, $a, b \in \mathbb{T}$ with $a < b$ and an interval $[a, b]_{\mathbb{T}}$ means the intersection of a real interval with the given time scale.

2. Preliminaries

We need here basic concepts of delta calculus. The results of delta calculus are adapted from monographs [7, 8].

For $t \in \mathbb{T}$, the *forward jump operator* $\sigma : \mathbb{T} \to \mathbb{T}$ is defined by

$$\sigma(t) := \inf\{s \in \mathbb{T} : s > t\}.$$

The mapping $\mu : \mathbb{T} \to \mathbb{R}_0^+ = [0, +\infty)$ such that $\mu(t) := \sigma(t) - t$ is called the *forward graininess function*. The *backward jump operator* $\rho : \mathbb{T} \to \mathbb{T}$ is defined by

$$\rho(t) := \sup\{s \in \mathbb{T} : s < t\}.$$

The mapping $\nu : \mathbb{T} \to \mathbb{R}_0^+ = [0, +\infty)$ such that $\nu(t) := t - \rho(t)$ is called the *backward graininess function*. If $\sigma(t) > t$, we say that t is *right-scattered*, while if $\rho(t) < t$, we say that t is *left-scattered*. Also, if $t < \sup \mathbb{T}$ and $\sigma(t) = t$, then t is called *right-dense*, and if $t > \inf \mathbb{T}$ and $\rho(t) = t$, then t is called *left-dense*. If $\mathbb{T}$ has a left-scattered maximum M, then $\mathbb{T}^k = \mathbb{T} - \{M\}$, otherwise $\mathbb{T}^k = \mathbb{T}$.

For a function $f : \mathbb{T} \to \mathbb{R}$, the delta derivative f^{Δ} is defined as follows:

Let $t \in \mathbb{T}^k$, if there exists $f^{\Delta}(t) \in \mathbb{R}$ such that for all $\epsilon > 0$, there exists a neighborhood U of t, such that

$$|f(\sigma(t)) - f(s) - f^{\Delta}(t)(\sigma(t) - s)| \leq \epsilon|\sigma(t) - s|,$$

for all $s \in U$, then f is said to be *delta differentiable* at t, and $f^{\Delta}(t)$ is called the *delta derivative* of f at t.

A function $f : \mathbb{T} \to \mathbb{R}$ is said to be *right-dense continuous* (*rd-continuous*), if it is continuous at each right-dense point and there exists a finite left limit at every left-dense point. The set of all rd-continuous functions is denoted by $C_{rd}(\mathbb{T}, \mathbb{R})$.

The next definition is given in [7, 8].

Definition 1. *A function $F:\mathbb{T}\to\mathbb{R}$ is called a delta antiderivative of $f:\mathbb{T}\to\mathbb{R}$, provided that $F^{\Delta}(t)=f(t)$ holds for all $t\in\mathbb{T}^k$. Then the delta integral of f is defined by*

$$\int_a^b f(t)\Delta t = F(b)-F(a).$$

The following results of nabla calculus are taken from [3, 7, 8].

If $\mathbb{T}$ has a right-scattered minimum m, then $\mathbb{T}_k=\mathbb{T}-\{m\}$, otherwise $\mathbb{T}_k=\mathbb{T}$. A function $f:\mathbb{T}_k\to\mathbb{R}$ is called *nabla differentiable* at $t\in\mathbb{T}_k$, with nabla derivative $f^{\nabla}(t)$, if there exists $f^{\nabla}(t)\in\mathbb{R}$ such that for any $\epsilon>0$, there exists a neighborhood V of t, such that

$$|f(\rho(t))-f(s)-f^{\nabla}(t)(\rho(t)-s)|\leq\epsilon|\rho(t)-s|,$$

for all $s\in V$.

A function $f:\mathbb{T}\to\mathbb{R}$ is said to be *left-dense continuous* (*ld-continuous*), provided it is continuous at all left-dense points in $\mathbb{T}$ and its right-sided limits exist (finite) at all right-dense points in $\mathbb{T}$. The set of all ld-continuous functions is denoted by $C_{ld}(\mathbb{T},\mathbb{R})$.

The next definition is given in [3, 7, 8].

Definition 2. *A function $G:\mathbb{T}\to\mathbb{R}$ is called a nabla antiderivative of $g:\mathbb{T}\to\mathbb{R}$, provided that $G^{\nabla}(t)=g(t)$ holds for all $t\in\mathbb{T}_k$. Then the nabla integral of g is defined by*

$$\int_a^b g(t)\nabla t = G(b)-G(a).$$

Now we present short introduction of diamond-α derivative as given in [1, 19].

Let $\mathbb{T}$ be a time scale and $f(t)$ be differentiable on $\mathbb{T}$ in the Δ and ∇ senses. For $t\in\mathbb{T}_k^k$, where $\mathbb{T}_k^k=\mathbb{T}^k\cap\mathbb{T}_k$, the *diamond-$\alpha$ dynamic derivative* $f^{\diamond_\alpha}(t)$ is defined by

$$f^{\diamond_\alpha}(t)=\alpha f^{\Delta}(t)+(1-\alpha)f^{\nabla}(t),\quad 0\leq\alpha\leq 1.$$

Thus f is diamond-α differentiable if and only if f is Δ and ∇ differentiable.

The diamond-α derivative reduces to the standard Δ-derivative for $\alpha=1$, or the standard ∇-derivative for $\alpha=0$. It represents a weighted dynamic derivative for $\alpha\in(0,1)$.

Theorem 4. [19] *Let $f,g:\mathbb{T}\to\mathbb{R}$ be diamond-α differentiable at $t\in\mathbb{T}$ and we write $f^{\sigma}(t)=f(\sigma(t))$, $g^{\sigma}(t)=g(\sigma(t))$, $f^{\rho}(t)=f(\rho(t))$ and $g^{\rho}(t)=g(\rho(t))$. Then*

(*i*) *$f\pm g:\mathbb{T}\to\mathbb{R}$ is diamond-α differentiable at $t\in\mathbb{T}$, with*

$$(f\pm g)^{\diamond_\alpha}(t)=f^{\diamond_\alpha}(t)\pm g^{\diamond_\alpha}(t).$$

(*ii*) *$fg:\mathbb{T}\to\mathbb{R}$ is diamond-α differentiable at $t\in\mathbb{T}$, with*

$$(fg)^{\diamond_\alpha}(t)=f^{\diamond_\alpha}(t)g(t)+\alpha f^{\sigma}(t)g^{\Delta}(t)+(1-\alpha)f^{\rho}(t)g^{\nabla}(t).$$

(*iii*) *For $g(t)g^{\sigma}(t)g^{\rho}(t)\neq 0$, $\frac{f}{g}:\mathbb{T}\to\mathbb{R}$ is diamond-α differentiable at $t\in\mathbb{T}$, with*

$$\left(\frac{f}{g}\right)^{\diamond_\alpha}(t)=\frac{f^{\diamond_\alpha}(t)g^{\sigma}(t)g^{\rho}(t)-\alpha f^{\sigma}(t)g^{\rho}(t)g^{\Delta}(t)-(1-\alpha)f^{\rho}(t)g^{\sigma}(t)g^{\nabla}(t)}{g(t)g^{\sigma}(t)g^{\rho}(t)}.$$

Definition 3. [19] *Let $a, t \in \mathbb{T}$ and $h : \mathbb{T} \to \mathbb{R}$. Then the diamond-$\alpha$ integral from a to t of h is defined by*

$$\int_a^t h(s) \diamond_\alpha s = \alpha \int_a^t h(s)\Delta s + (1-\alpha) \int_a^t h(s)\nabla s, \quad 0 \le \alpha \le 1,$$

provided that there exist delta and nabla integrals of h on $\mathbb{T}$.

Theorem 5. [19] *Let $a, b, t \in \mathbb{T}$, $c \in \mathbb{R}$. Assume that $f(s)$ and $g(s)$ are $\diamond_\alpha$-integrable functions on $[a, b]_{\mathbb{T}}$. Then*

(i) $\int_a^t [f(s) \pm g(s)] \diamond_\alpha s = \int_a^t f(s) \diamond_\alpha s \pm \int_a^t g(s) \diamond_\alpha s$;

(ii) $\int_a^t cf(s) \diamond_\alpha s = c \int_a^t f(s) \diamond_\alpha s$;

(iii) $\int_a^t f(s) \diamond_\alpha s = -\int_t^a f(s) \diamond_\alpha s$;

(iv) $\int_a^t f(s) \diamond_\alpha s = \int_a^b f(s) \diamond_\alpha s + \int_b^t f(s) \diamond_\alpha s$;

(v) $\int_a^a f(s) \diamond_\alpha s = 0$.

We need the following results.

Definition 4. [9] *A function $f : \mathbb{T} \to \mathbb{R}$ is called convex on $I_{\mathbb{T}} = I \cap \mathbb{T}$, where I is an interval of $\mathbb{R}$ (open or closed), if*

$$f(\lambda t + (1-\lambda)s) \le \lambda f(t) + (1-\lambda) f(s), \tag{2.1}$$

for all $t, s \in I_{\mathbb{T}}$ and all $\lambda \in [0, 1]$ such that $\lambda t + (1-\lambda)s \in I_{\mathbb{T}}$.

The function f is strictly convex on $I_{\mathbb{T}}$ if the inequality (2.1) is strict for distinct $t, s \in I_{\mathbb{T}}$ and $\lambda \in (0, 1)$.

The function f is concave (respectively, strictly concave) on $I_{\mathbb{T}}$, if $-f$ is convex (respectively, strictly convex).

Theorem 6. [1] *Let $a, b \in \mathbb{T}$ and $c, d \in \mathbb{R}$. Suppose that $g \in C([a, b]_{\mathbb{T}}, (c, d))$ and $h \in C([a, b]_{\mathbb{T}}, \mathbb{R})$ with $\int_a^b |h(s)| \diamond_\alpha s > 0$. If $\Phi \in C((c, d), \mathbb{R})$ is convex, then generalized Jensen's Inequality is*

$$\Phi\left(\frac{\displaystyle\int_a^b |h(s)| g(s) \diamond_\alpha s}{\displaystyle\int_a^b |h(s)| \diamond_\alpha s}\right) \le \frac{\displaystyle\int_a^b |h(s)| \Phi(g(s)) \diamond_\alpha s}{\displaystyle\int_a^b |h(s)| \diamond_\alpha s}. \tag{2.2}$$

If Φ is strictly convex, then the inequality $\le$ can be replaced by $<$.

Example 1. [1] One of the three most popular examples of calculus on time scales is quantum calculus, i. e., $q^{\mathbb{N}_0} = \{q^n : n \in \mathbb{N}_0\}$, where $\mathbb{N}_0$ is the set of nonnegative integers.

If we set $\mathbb{T} = q^{\mathbb{N}_0}$ for $q > 1$ and $m < n$, then

$$\int_{q^m}^{q^n} f(x) \diamond_\alpha x = (q-1) \sum_{i=m}^{n-1} q^i [\alpha f(q^i) + (1-\alpha) f(q^{i+1})],$$

for $m, n \in \mathbb{N}_0$.

3. Main Results

In order to present our main results, first we present an extension of Radon's Inequality via time scales.

Theorem 7. *Let $w, f, g \in C([a,b]_{\mathbb{T}}, \mathbb{R})$ be $\diamond_\alpha$-integrable functions, where $w(x)$, $g(x) \neq 0$, $\forall x \in [a,b]_{\mathbb{T}}$. If $\beta \geq \gamma \geq 0$, then*

$$\left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\gamma-\beta} \frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{\gamma}} \leq \int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\gamma}} \diamond_\alpha x. \tag{3.1}$$

Equality holds in (3.1), when $f(x) \equiv g(x) \equiv c$, where c is a nonzero real constant.

P r o o f. The right hand side of (3.1) can be written as

$$\int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\gamma}} \diamond_\alpha x = \int_a^b |w(x)||g(x)| \diamond_\alpha x \times \int_a^b \frac{|w(x)||g(x)|(\Psi(x))^{\gamma+1}}{\int_a^b |w(x)||g(x)| \diamond_\alpha x} \diamond_\alpha x, \tag{3.2}$$

where

$$\Psi(x) = \frac{|f(x)|^{\frac{\beta+1}{\gamma+1}}}{|g(x)|}.$$

The function $\Phi : [0,\infty) \to [0,\infty)$ defined by $\Phi(x) = x^{\gamma+1}$ is convex for $x \in [0,\infty)$, so applying generalized Jensen's Inequality given in (2.2), we have

$$\left(\int_a^b \frac{|w(x)||g(x)|\Psi(x)}{\int_a^b |w(x)||g(x)| \diamond_\alpha x} \diamond_\alpha x\right)^{\gamma+1} \leq \int_a^b \frac{|w(x)||g(x)|(\Psi(x))^{\gamma+1}}{\int_a^b |w(x)||g(x)| \diamond_\alpha x} \diamond_\alpha x. \tag{3.3}$$

Then (3.3) can be written as

$$\left(\int_a^b \frac{|w(x)||f(x)|^{\frac{\beta+1}{\gamma+1}}}{\int_a^b |w(x)||g(x)| \diamond_\alpha x} \diamond_\alpha x\right)^{\gamma+1} \leq \int_a^b \frac{|w(x)||g(x)|\frac{|f(x)|^{\beta+1}}{|g(x)|^{\gamma+1}}}{\int_a^b |w(x)||g(x)| \diamond_\alpha x} \diamond_\alpha x. \tag{3.4}$$

Now the function $\Phi : [0,\infty) \to [0,\infty)$ defined by $\Phi(x) = x^{\frac{\beta+1}{\gamma+1}}$ is convex for $x \in [0,\infty)$, so applying generalized Jensen's Inequality, we have

$$\left(\int_a^b \frac{|w(x)||f(x)| \diamond_\alpha x}{\int_a^b |w(x)| \diamond_\alpha x}\right)^{\frac{\beta+1}{\gamma+1}} \leq \int_a^b \frac{|w(x)||f(x)|^{\frac{\beta+1}{\gamma+1}} \diamond_\alpha x}{\int_a^b |w(x)| \diamond_\alpha x}. \tag{3.5}$$

Now (3.5) becomes

$$\left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\frac{\gamma-\beta}{\gamma+1}} \left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\frac{\beta+1}{\gamma+1}} \leq \int_a^b |w(x)||f(x)|^{\frac{\beta+1}{\gamma+1}} \diamond_\alpha x. \tag{3.6}$$

From (3.2), (3.4) and (3.6), we get (3.1).

Clearly equality holds in (3.1), when $f(x) \equiv g(x) \equiv c$, where c is a nonzero real constant, which completes the proof. $\square$

Remark 1. If we set $\alpha = 1$, $\mathbb{T} = \mathbb{Z}$, $w(x) = 1$, $\beta = \gamma = 1$, $f(k) = x_k \in \mathbb{R}$ and $g(k) = y_k \in (0, \infty)$ for $k \in \{1, 2, ..., n\}$, $n \in \mathbb{N}$, then (3.1) reduces to (1.1).

Remark 2. If we set $\alpha = 1$, $\mathbb{T} = \mathbb{Z}$, $w(x) = 1$, $\beta = \gamma \geq 0$, $f(k) = x_k \in [0, \infty)$ and $g(k) = y_k \in (0, \infty)$ for $k \in \{1, 2, ..., n\}$, $n \in \mathbb{N}$, then (3.1) reduces to (1.2).

Remark 3. If we set $\alpha = 1$, $\mathbb{T} = \mathbb{Z}$, $w(x) = 1$, $\beta \geq \gamma \geq 0$, $f(k) = x_k \in [0, \infty)$ and $g(k) = y_k \in (0, \infty)$ for $k \in \{1, 2, ..., n\}$, $n \in \mathbb{N}$, then discrete version of (3.1) reduces to (1.3).

Example 2. When $\mathbb{T} = \mathbb{R}$, then continuous version of (3.1) can be written as

$$\left(\int_a^b |w(x)|dx\right)^{\gamma-\beta} \frac{\left(\int_a^b |w(x)||f(x)|dx\right)^{\beta+1}}{\left(\int_a^b |w(x)||g(x)|dx\right)^{\gamma}} \leq \int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\gamma}}dx.$$

If we set $[a, b]_{\mathbb{T}} = [q^m, q^n]_{q^{\mathbb{N}_0}}$ for $q > 1$ and $m < n$, where $m, n \in \mathbb{N}_0$ and $\mathbb{N}_0$ is the set of nonnegative integers, then

$$\left[\sum_{i=m}^{n-1} q^i \left\{\alpha|w(q^i)|+(1-\alpha)|w(q^{i+1})|\right\}\right]^{\gamma-\beta} \times \frac{\left[\sum_{i=m}^{n-1} q^i \left\{\alpha|w(q^i)||f(q^i)|+(1-\alpha)|w(q^{i+1})||f(q^{i+1})|\right\}\right]^{\beta+1}}{\left[\sum_{i=m}^{n-1} q^i \left\{\alpha|w(q^i)||g(q^i)|+(1-\alpha)|w(q^{i+1})||g(q^{i+1})|\right\}\right]^{\gamma}}$$
$$\leq \sum_{i=m}^{n-1} q^i \left\{\alpha\frac{|w(q^i)||f(q^i)|^{\beta+1}}{|g(q^i)|^{\gamma}} + (1-\alpha)\frac{|w(q^{i+1})||f(q^{i+1})|^{\beta+1}}{|g(q^{i+1})|^{\gamma}}\right\}.$$

Corollary 1. *Let $w, f, g \in C([a, b]_{\mathbb{T}}, \mathbb{R} - \{0\})$ be $\diamond_\alpha$-integrable functions. If $\gamma \leq \beta \leq -1$, then*

$$\left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\gamma-\beta} \frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{\gamma}} \leq \int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\gamma}} \diamond_\alpha x. \tag{3.7}$$

Equality holds in (3.7), when $f(x) \equiv g(x) \equiv c$, where c is a nonzero real constant.

P r o o f. For $\beta \leq -1$, $\gamma \leq -1$, the inequalities $-\gamma \geq -\beta$, $-\gamma \geq 1$, $-\beta \geq 1$ hold. Taking into account inequality (3.1), we obtain

$$\int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\gamma}} \diamond_\alpha x = \int_a^b \frac{|w(x)||g(x)|^{-\gamma}}{|f(x)|^{-\beta-1}} \diamond_\alpha x$$
$$\geq \left(\int_a^b |w(x)| \diamond_\alpha x\right)^{-\beta+\gamma} \frac{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{-\gamma}}{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{-\beta-1}}$$
$$= \left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\gamma-\beta} \frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{\gamma}},$$

thus inequality (3.7) holds. Clearly the equality holds in (3.7), when $f(x) \equiv g(x) \equiv c$, where c is a nonzero real constant. □

Corollary 2. *Let $w, f, g \in C([a,b]_{\mathbb{T}}, \mathbb{R})$ be $\diamond_\alpha$-integrable functions, where $w(x)$, $g(x) \neq 0$, $\forall x \in [a,b]_{\mathbb{T}}$. If $\beta \geq 0$, then*

$$\frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{\beta}} \leq \int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\beta}} \diamond_\alpha x. \tag{3.8}$$

Equality holds in (3.8), when $f(x) = cg(x)$, where c is a real constant.

P r o o f. If we put $\beta = \gamma$ in (3.1), then we get (3.8), which is Radon's Inequality on dynamic time scales. Clearly the equality holds in (3.8), if $f(x) = cg(x)$, where c is a real constant. □

Corollary 3. *Let $w, f, g \in C([a,b]_{\mathbb{T}}, \mathbb{R} - \{0\})$ be $\diamond_\alpha$-integrable functions. If $\beta \leq -1$, then*

$$\frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{\beta}} \leq \int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\beta}} \diamond_\alpha x. \tag{3.9}$$

Equality holds in (3.9), when $f(x) = cg(x)$, where c is a nonzero real constant.

P r o o f. By applying inequality (3.8) for $\beta \leq -1$, we obtain

$$\int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{|g(x)|^{\beta}} \diamond_\alpha x = \int_a^b \frac{|w(x)||g(x)|^{-\beta}}{|f(x)|^{-\beta-1}} \diamond_\alpha x \geq \frac{\left(\int_a^b |w(x)||g(x)| \diamond_\alpha x\right)^{-\beta}}{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{-\beta-1}},$$

thus inequality (3.9) holds. Clearly the equality holds in (3.9), if $f(x) = cg(x)$, where c is a nonzero real constant. □

Corollary 4. *Let $w, f, g \in C([a,b]_{\mathbb{T}}, \mathbb{R} - \{0\})$ be $\diamond_\alpha$-integrable functions. If $\beta > 0$ or $\beta \leq -1$, then*

$$\frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left(\int_a^b |w(x)||f(x)||g(x)|^{\frac{1}{\beta}} \diamond_\alpha x\right)^{\beta}} \leq \int_a^b \frac{|w(x)||f(x)|}{|g(x)|} \diamond_\alpha x. \tag{3.10}$$

P r o o f. Replace $|g(x)|$ by $|f(x)||g(x)|^{\frac{1}{\beta}}$ in (3.8) and (3.9), then inequality (3.10) is obtained. □

Corollary 5. *Let $w, f \in C([a,b]_{\mathbb{T}}, \mathbb{R})$ be $\diamond_\alpha$-integrable functions, where $w(x) \neq 0$, $\forall x \in [a,b]_{\mathbb{T}}$. If $\eta_2 \geq \eta_1 > 0$, then*

$$\left(\frac{\int_a^b |w(x)||f(x)|^{\eta_1} \diamond_\alpha x}{\int_a^b |w(x)| \diamond_\alpha x}\right)^{\frac{1}{\eta_1}} \leq \left(\frac{\int_a^b |w(x)||f(x)|^{\eta_2} \diamond_\alpha x}{\int_a^b |w(x)| \diamond_\alpha x}\right)^{\frac{1}{\eta_2}}. \tag{3.11}$$

P r o o f. Take $\beta \geq 0$, $1+\beta = \frac{\eta_2}{\eta_1} \geq 1$ and $g(x)=1$, then (3.8) becomes

$$\frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\frac{\eta_2}{\eta_1}}}{\left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\frac{\eta_2}{\eta_1}-1}} \leq \int_a^b |w(x)||f(x)|^{\frac{\eta_2}{\eta_1}} \diamond_\alpha x. \tag{3.12}$$

Dividing (3.12) by $\int_a^b |w(x)| \diamond_\alpha x$, replacing $|f(x)|$ by $|f(x)|^{\eta_1}$ and taking power $1/\eta_2 > 0$, we get the inequality (3.11), which is known as the dynamic weighted power mean inequality on time scales. □

Remark 4. Set $\beta \geq 0$, $1+\beta = \eta_2/\eta_1 \geq 1$, $g(x)=1$ and $\int_a^b |w(x)| \diamond_\alpha x = 1$, then (3.11) becomes

$$\left(\int_a^b |w(x)||f(x)|^{\eta_1} \diamond_\alpha x\right)^{\frac{1}{\eta_1}} \leq \left(\int_a^b |w(x)||f(x)|^{\eta_2} \diamond_\alpha x\right)^{\frac{1}{\eta_2}}. \tag{3.13}$$

Inequality given in (3.13) is called Schlömilch's Inequality on time scales.

Let $w, f \in C_{rd}([a,b]_{\mathbb{T}}, [0,\infty))$, then for $\alpha = 1$, inequality (3.13) takes the form

$$\left(\int_a^b w(x) f^{\eta_1}(x) \Delta x\right)^{\frac{1}{\eta_1}} \leq \left(\int_a^b w(x) f^{\eta_2}(x) \Delta x\right)^{\frac{1}{\eta_2}},$$

as given in [12, Lemma A].

Let $w, f \in C([a,b]_{\mathbb{T}}, [0,\infty))$, then inequality (3.13) takes the form

$$\left(\int_a^b w(x) f^{\eta_1}(x) \diamond_\alpha x\right)^{\frac{1}{\eta_1}} \leq \left(\int_a^b w(x) f^{\eta_2}(x) \diamond_\alpha x\right)^{\frac{1}{\eta_2}},$$

as given in [20, Lemma 3.4].

Now we present generalized Nesbitt's Inequality on dynamic time scale calculus.

Theorem 8. *Let $w, f \in C([a,b]_{\mathbb{T}}, \mathbb{R} - \{0\})$ be $\diamond_\alpha$-integrable functions, $c, d \in \mathbb{R}$ and*

$$c \int_a^b |w(x)||f(x)| \diamond_\alpha x - d|f(x)| > 0,$$

where $x \in [a,b]_{\mathbb{T}}$. If $\beta \geq \gamma \geq 0$, then

$$\begin{aligned} &\left(\int_a^b |w(x)| \diamond_\alpha x\right)^{2\gamma-\beta} \frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta-2\gamma+1}}{\left(c\int_a^b |w(x)| \diamond_\alpha x - d\right)^{\gamma}} \\ &\leq \int_a^b \frac{|w(x)||f(x)|^{\beta-\gamma+1}}{\left(c\int_a^b |w(x)||f(x)| \diamond_\alpha x - d|f(x)|\right)^{\gamma}} \diamond_\alpha x. \end{aligned} \tag{3.14}$$

P r o o f. By using inequality (3.1), we have that

$$\begin{aligned}
&\int_a^b \frac{|w(x)||f(x)|^{\beta-\gamma+1}}{\left(c\int_a^b |w(x)||f(x)| \diamond_\alpha x - d|f(x)|\right)^\gamma} \diamond_\alpha x \\
&= \int_a^b \frac{|w(x)||f(x)|^{\beta+1}}{\left(c|f(x)|\int_a^b |w(x)||f(x)| \diamond_\alpha x - d|f(x)|^2\right)^\gamma} \diamond_\alpha x \\
&\geq \left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\gamma-\beta} \frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left[c\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^2 - d\int_a^b |w(x)||f(x)|^2 \diamond_\alpha x\right]^\gamma}.
\end{aligned}$$

Taking Jensen's Inequality into account, we have

$$\left(\int_a^b \frac{|w(x)||f(x)| \diamond_\alpha x}{\int_a^b |w(x)| \diamond_\alpha x}\right)^2 \leq \int_a^b \frac{|w(x)||f(x)|^2 \diamond_\alpha x}{\int_a^b |w(x)| \diamond_\alpha x}. \tag{3.15}$$

Then, using inequality (3.15), we have that

$$\begin{aligned}
&c\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^2 - d\int_a^b |w(x)||f(x)|^2 \diamond_\alpha x \\
&\leq c\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^2 - \frac{d}{\int_a^b |w(x)| \diamond_\alpha x}\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^2 \\
&= \left(\frac{c\int_a^b |w(x)| \diamond_\alpha x - d}{\int_a^b |w(x)| \diamond_\alpha x}\right)\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^2.
\end{aligned}$$

And then

$$\begin{aligned}
&\frac{\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta+1}}{\left[c\left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^2 - d\int_a^b |w(x)||f(x)|^2 \diamond_\alpha x\right]^\gamma} \\
&\geq \left(\frac{\int_a^b |w(x)| \diamond_\alpha x}{c\int_a^b |w(x)| \diamond_\alpha x - d}\right)^\gamma \left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta-2\gamma+1}.
\end{aligned}$$

So,

$$\left(\frac{\displaystyle\int_a^b |w(x)| \diamond_\alpha x}{c\int_a^b |w(x)| \diamond_\alpha x - d}\right)^\gamma \left(\int_a^b |w(x)||f(x)| \diamond_\alpha x\right)^{\beta-2\gamma+1}$$
$$\leq \left(\int_a^b |w(x)| \diamond_\alpha x\right)^{\beta-\gamma} \int_a^b \frac{|w(x)||f(x)|^{\beta-\gamma+1}}{\left(c\int_a^b |w(x)||f(x)| \diamond_\alpha x - d|f(x)|\right)^\gamma} \diamond_\alpha x. \tag{3.16}$$

Clearly (3.14) holds true from (3.16). □

Remark 5. If we set $\alpha = 1$, $\mathbb{T} = \mathbb{Z}$, $w(x) = 1$, $\beta = \gamma = 1$ and $f(k) = x_k \in (0, \infty)$ for $k \in \{1, 2, ..., n\}$, $n \in \mathbb{N} - \{1\}$, then discrete version of (3.14) reduces to

$$\frac{n}{cn - d} \leq \sum_{k=1}^{n} \frac{x_k}{cX_n - dx_k}, \tag{3.17}$$

where $X_n = x_1 + x_2 + ... + x_n$.

Inequality (3.17) is called generalized Nesbitt's Inequality as given in [10].

Further if we set $n = 3$ and $c = d$, where $c, d \in (0, \infty)$, then (3.17) takes the form

$$\frac{3}{2} \leq \frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_1} + \frac{x_3}{x_1 + x_2}, \tag{3.18}$$

where $x_1, x_2, x_3 > 0$. Inequality (3.18) is called Nesbitt's Inequality as given in [14].

Remark 6. If we set $\alpha = 1$, then we get delta versions of dynamic inequalities and if we set $\alpha = 0$, then we get nabla versions of dynamic inequalities presented in this article.

Further if we put $\mathbb{T} = \mathbb{Z}$, then we get discrete versions of dynamic inequalities and if we put $\mathbb{T} = \mathbb{R}$, then we get continuous versions of dynamic inequalities presented in this article.

4. Conclusion and Future Work

In this research article, we have presented dynamic inequalities on diamond-α calculus such as Radon's Inequality, Bergström's Inequality, the weighted power mean inequality, Schlömilch's Inequality and Nesbitt's Inequality.

We can generalize dynamic inequalities using functional generalization as given in [20]. We can present dynamic inequalities on fractional calculus as given in [16] and on quantum calculus. We can develop dynamic inequalities using fractional Riemann-Liouville integral on time scale calculus in a similar fashion as given in [2] and [17]. Similarly we can generalize dynamic inequalities of this article using time scales fractional derivative as given in [2]. It will be interesting to present dynamic inequalities in two or more dimensions.

Recently it has found that many dynamic inequalities such as Radon's Inequality, the weighted power mean inequality, Schlömilch's Inequality, Rogers-Hölder's Inequality and Bernoulli's Inequality are equivalent on time scales as given in [18], so we can find more equivalent dynamic inequalities on time scales.

REFERENCES

1. Agarwal R. P., O'Regan D., Saker S. H. *Dynamic Inequalities on Time Scales.* Springer International Publishing, Cham, Switzerland, 2014. DOI: 10.1007/978-3-319-11002-8

2. Anastassiou G. A. Integral operator inequalities on time scales. *Intern. J. Difference Equ.*, 2012. Vol. 7, No. 2. P. 111–137.
3. Anderson D., Bullock J., Erbe L., Peterson A., Tran H. Nabla dynamic equations on time scales. *Pan-American Math. J.*, 2003. Vol. 13, No. 1. P. 1–48.
4. Beckenbach E. F., Bellman R. *Inequalities.* Springer, Berlin, Göttingen and Heidelberg, 1961. DOI: 10.1007/978-3-642-64971-4
5. Bellman R. Notes on matrix theory-IV (An inequality due to Bergström). *Amer. Math. Monthly*, 1955. Vol. 62. P. 172–173.
6. Bergström H. A triangle inequality for matrices. *Den Elfte Skandinaviske Matematikerkongress*, 1949, *Trondheim, Johan Grundt Tanums Forlag, Oslo*, 1952. P. 264–267.
7. Bohner M., Peterson A. *Dynamic Equations on Time Scales.* Birkhäuser Boston, Inc., Boston, MA, 2001. DOI: 10.1007/978-1-4612-0201-1
8. Bohner M., Peterson A. *Advances in Dynamic Equations on Time Scales.* Birkhäuser Boston, Boston, MA, 2003. DOI: 10.1007/978-0-8176-8230-9
9. Dinu C. Convex functions on time scales. *Annals Univ. of Craiova, Math. Comp. Sci. Ser.*, 2008. Vol. 35. P. 87–96.
10. Bătineţu-Giurgiu D. M., Mărghidanu D., Pop O. T. A new generalization of Radon's Inequality and applications. *Creative Math & Inf*, 2011. Vol. 20, No. 2. P. 111–116.
11. Hilger S. *Ein Maßkettenkalkül mit Anwendung auf Zentrumsmannigfaltigkeiten.* PhD thesis, Universität Würzburg, 1988.
12. Hong C. H., Yeh C. C. Rogers-Hölder's Inequality on time scales. *Intern. J. Pure Appl. Math.*, 2006. Vol. 29, No. 3. P. 289–309.
13. Mitrinović D. S. *Analytic Inequalities.* Springer-Verlag, Berlin, 1970. DOI: 10.1007/978-3-642-99970-3
14. Nesbitt A. M. Problem 15114. *Educational Times*, 1903. Vol. 3. P. 37–38.
15. Radon J. Theorie und Anwendungen der absolut additiven Mengenfunktionen. *Sitzungsber. Acad. Wissen. Wien*, 1913. Vol. 122. P. 1295–1438.
16. Sahir M. J. S. Dynamic inequalities for convex functions harmonized on time scales. *J. Appl. Math. Phys.*, 2017. Vol. 5. P. 2360–2370. DOI: 10.4236/jamp.2017.512193
17. Sahir M. J. S. Fractional dynamic inequalities harmonized on time scales. *Cogent Math. Stat.*, 2018. Vol. 5. P. 1–7. DOI: 10.1080/23311835.2018.1438030
18. Sahir M. J. S. Hybridization of classical inequalities with equivalent dynamic inequalities on time scale calculus. *The Teaching of Mathematics*, 2018. Vol. 21, No. 1. P. 38–52.
19. Sheng Q., Fadag M., Henderson J., Davis J. M. An exploration of combined dynamic derivatives on time scales and their applications. *Nonlinear Anal. Real World Appl.*, 2006. Vol. 7, No. 3. P. 395–413. DOI: 10.1016/j.nonrwa.2005.03.008
20. Tian J. F. Ha M. H. Extensions of Hölder-type inequalities on time scales and their applications. *J. Nonlinear Sci. Appl.*, 2017. Vol. 10, No. 3. P. 937–953. DOI: 10.22436/jnsa.010.03.07

URAL MATHEMATICAL JOURNAL, Vol. 4, No. 2, 2018, pp. 99–110

DOI: 10.15826/umj.2018.2.011

D_2-SYNCHRONIZATION IN NONDETERMINISTIC AUTOMATA[1]

Hanan Shabana

Institute of Natural Sciences and Mathematics,

Ural Federal University, 51 Lenin aven., Ekaterinburg, Russia, 620000

Faculty of Electronic Engineering, Menoufia University, Egypt

hananshabana22@gmail.com

Abstract: We approach the problem of computing a D_2-synchronizing word of minimum length for a given nondeterministic automaton via its encoding as an instance of SAT and invoking a SAT solver. In addition, we report some of the experimental results obtained when we had tested our method on randomly generated automata and certain benchmarks.

Keywords: Nondeterministic automata, Synchronizing word, SAT solver

Introduction

A *nondeterministic finite automaton* (NFA) is a triple $\mathscr{A} = (Q, \Sigma, \delta)$, where Q is a finite non-empty set of *states*, Σ is a finite non-empty set of input symbols, and δ is a map $Q \times \Sigma \to \mathcal{P}(Q)$, where $\mathcal{P}(Q)$ is the power set of Q. The map δ is called the *transition function* of $\mathscr{A}$; it describes the action of symbols in Σ at states in Q. As usual, we represent the NFA $\mathscr{A}$ by the labeled digraph with the vertex set Q, the label alphabet Σ, and the set of labeled edges

$$\{q \xrightarrow{s} q' \mid q, q' \in Q,\ s \in \Sigma,\ q' \in \delta(q, s)\}.$$

A *word* over Σ is a finite (maybe, empty) sequence of symbols from Σ. The set of all words over Σ including the empty word is denoted by Σ^*. If $w = a_1 \cdots a_\ell$ with $a_1, \ldots, a_\ell \in \Sigma$ is a non-empty word over Σ, the number ℓ is said to be the *length* of w and is denoted by $|w|$. The length of the empty word is defined to be 0. The set of all words of a given length ℓ over Σ is denoted by Σ^ℓ.

For every NFA $\mathscr{A} = \langle Q, \Sigma, \delta \rangle$, we extend the function δ to a function $\mathcal{P}(Q) \times \Sigma^* \to \mathcal{P}(Q)$ (still denoted by δ) by induction on the length of $w \in \Sigma^*$. If $|w| = 0$, that is, w is the empty word, then, for each $X \subseteq Q$, we let $\delta(X, w) := X$. If $|w| > 0$, we represent w as $w = sw'$ with $w' \in \Sigma^*$ and $s \in \Sigma$ and, for each $X \subseteq Q$, let $\delta(X, w) := \bigcup_{q \in X} \delta(\delta(q, s), w')$ (the right hand side of the latter equality is defined by the induction assumption since $|w'| < |w|$). To lighten the notation, we write $q.w$ for $\delta(q, w)$ and $X.w$ for $\delta(X, w)$ whenever we deal with a fixed automaton.

The present note is a follow-up of the paper [12] by Volkov and the present author. We briefly recall the problem approached in [12] and, in parallel, introduce the problem that we tackle here. We are interested in *synchronization* of finite automata. The basic idea of *synchronization* is as follows: for a given automaton, we look for a sequence of input signals that allows us to predict the behaviour of the automaton after consuming these signals, no matter at which state the automaton was at the beginning. This input is called a *synchronizing word*, and if an automaton possesses such a word, it is called *synchronizing*.

[1]Supported by the Competitiveness Enhancement Program of Ural Federal University.

The above informal idea of synchronization is fairly easy to formalize for deterministic automata but for NFAs it admits several non-equivalent formalizations. We are not going to survey all formalizations that appear in the literature and restrict ourselves to the two following versions of synchronization, both originating in [7].

Let $\mathscr{A} = (Q, \Sigma, \delta)$ be an NFA. A word $w \in \Sigma^*$ is said to be D_3-*synchronizing* if $\bigcap_{q\in Q} q.w \neq \varnothing$. In terms of the labeled digraph representing $\mathscr{A}$, this condition amounts to saying that for each $q \in Q$, there exists a directed path, whose consecutive labels form the word w, that starts at q and terminates at a certain common state, independent of q. Observe that this definition implies that the action of any D_3-synchronizing word must be defined at every state of $\mathscr{A}$. A NFA is called D_3-*synchronizing* if it admits a D_3-synchronizing word.

A word $w \in \Sigma^*$ is said to be D_2-*synchronizing* for $\mathscr{A} = (Q, \Sigma, \delta)$ if $q.w = q'.w$ for all $q, q' \in Q$. To understand the 'physical meaning' of this concept, imagine a big quantity of identical NFAs which get the same input sequence and work on it in parallel. If the sequence constitutes a D_2-synchronizing word, then after consuming the input, the NFAs will demonstrate identical (that is, synchronous) behaviour, even though originally they all might have been in different states that were unknown to us.

In contrast to the condition $\bigcap_{q\in Q} q.w \neq \varnothing$, the equality $q.w = q'.w$ does not imply that the action of w must be everywhere defined. However, the equality ensures that if a D_2-synchronizing word is undefined at some state, the word must be nowhere defined. Thus, a D_2-synchronizing word is either nowhere or everywhere defined; in the latter case, it is easy to see that the word is also D_3-synchronizing. (The converse is not true: a D_3-synchronizing word can fail to be D_2-synchronizing.) A NFA is called D_2-*synchronizing* if it has a D_2-synchronizing word.

We mention that both D_3- and D_2-synchronization get very transparent meanings within a standard matrix representation of NFAs. In this representation, an NFA $\mathscr{A} = (Q, \Sigma, \delta)$ becomes a collection of $|\Sigma|$ Boolean $Q \times Q$-matrices where each input symbol $s \in \Sigma$ is encoded by a matrix $M(s)$ whose (q, q')-entry is 1 if $q' \in \delta(q, s)$ and 0 otherwise. It is not hard to check that the automaton $\mathscr{A}$ is D_3-synchronizing if and only if some product of the matrices $M(s)$, $s \in \Sigma$, has a column consisting entirely of 1s, and $\mathscr{A}$ is D_2-synchronizing if and only if in some product of the matrices $M(s)$, $s \in \Sigma$, every column consists either entirely of 0s or entirely of 1s.

The problems of determining whether or not a given NFA is either D_3- or D_2-synchronizing are known to be PSPACE-complete and there is no polynomial in n upper bound on the length of D_3- or D_2-synchronizing words for NFAs with n states. (These facts follow from results found by Rystsov in the early 1980s [10, 11] and later rediscovered (and strengthened) by Martyugin [9].) Thus, given an NFA, finding a D_3- or D_2-synchronizing word of minimum length for it is computationally hard. In [12] the author and Volkov have approached the problem of finding a D_3-synchronizing word of minimum length for a given NFA via the *SAT-solver method*. The method of treating computationally hard problems consists in encoding them as instances of the Boolean satisfiability problem (SAT) that are then fed to a SAT-solver, that is, a specialized program designed to solve instances of SAT. Modern SAT solvers are extremely powerful: they can solve instances with hundreds of thousands of variables and millions of clauses within a few minutes. Therefore the SAT-solver method has a very wide range of applications, see [5] for a survey. In particular, the method has been successfully invoked for studying synchronization in deterministic automata, see [6, 13]. Our results in [12] have demonstrated that the SAT-solver method can also be applied in the realm of NFAs. Here we extend the approach to the case of D_2-synchronization.

The paper is organized as follows. Sect. 1 describes the encoding reducing our problem to SAT. Sect. 2 presents the main algorithm, outlines the settings of our experiments and gives samples of our experimental results. Sect. 3 collects a few final remarks, including a discussion of the future work in the direction of the present paper.

1. Encoding to SAT

We start with a precise formulation of the problem which we are going to study here.

D2W (the existence of a D_2-synchronizing word of a given length):
INPUT: A NFA $\mathscr{A}$ with two input symbols and a positive integer ℓ.
OUTPUT: YES if $\mathscr{A}$ has a D_2-synchronizing word of length ℓ; NO otherwise.

In [12] the present author and Volkov have considered the problem D3W that has exactly the same instances as D2W but asks whether or not $\mathscr{A}$ has a D_3-synchronizing word of length ℓ. For both D2W and D3W, the integer ℓ is assumed to be given in unary; as explained in [12, Sect. 2], with ℓ given in binary, it is not feasible to expect the existence of a polynomial reduction from D3W to SAT, and the very same argument applies to D2W.

It is fair to say that our encoding of D2W has been obtained as a modification of the encoding of D3W suggested in [12]. However, restricting here to the "new" part of the encoding only would make the present paper difficult to follow without looking at [12] at every single step of the way. Therefore, we have preferred to describe our encoding in a self-contained manner, even though this causes a few overlaps with [12].

Recall that an instance of SAT is a pair (V, C), where V is a set of Boolean variables and C is a collection of clauses over V. (A *clause* over V is a disjunction of literals and a *literal* is either a variable in V or the negation of a variable in V.) The answer to an instance (V, C) is YES if (V, C) has a *satisfying assignment* (i.e., a truth assignment on V that satisfies C) and NO otherwise. We aim to construct a polynomial reduction of D2W to SAT. For this, we have to find two polynomials $v(x, y)$ and $c(x, y)$ (preferably of low degrees in x and y) with the following property: given an arbitrary instance $(\mathscr{A}, \ell)$ of D2W, where $\mathscr{A} = (Q, \Sigma, \delta)$ is an NFA with two input symbols, we are able to produce an instance (V, C) of SAT such that the answer to $(\mathscr{A}, \ell)$ is YES if and only if so is the answer to (V, C), while $|V| \leq v(|Q|, \ell)$ and $|C| \leq c(|Q|, \ell)$.

Throughout our encoding, we let $\Sigma := \{0, 1\}$ and $Q := \{q_0, \dots, q_{n-1}\}$. For a state $q \in Q$, we use the expressions $P_0(q)$ and $P_1(q)$ to denote the sets of all preimages of q under the actions of the input symbols 0 and 1 respectively; that is, if a is either of the two symbols, then

$$P_a(q) := \{p \in Q \mid q \in p.a\}.$$

First we define the set V of variables. We need two sorts of variables: *letter variables* and *token variables*.

The letter variables are $x_1, \dots, x_\ell$. The variable x_t, $1 \leq t \leq \ell$, plays the role of an indicator for the t-th symbols a_t in the input word $w := a_1 \cdots a_\ell \in \Sigma^\ell$: the value of x_t is 1 if and only if $a_t = 1$.

The token variables are y_{ij}^t where $i, j = 0, \dots, n-1$ and $t = 0, 1, 2, \dots, \ell$. To explain the role of these variables, we use a solitaire-like game Γ on the labeled digraph representing the NFA $\mathscr{A}$. In the initial position of Γ, each state $q_i \in Q$ holds exactly one token denoted $\mathbf{i}$. In the course of the game, tokens migrate and may multiply or disappear according to the previous position of the game and the action of the player. Namely, at each move an input symbol $a \in \Sigma$ is chosen. Then for each state $q \in Q$ such that $q.a \neq \varnothing$, all tokens that were held by q slide along the edges labeled a to all states in the set $q.a$. (If $|q.a| > 1$, then every token held by q gives rise to $|q.a|$ identical tokens, one for each state in $q.a$.) If $q.a = \varnothing$, then all tokens that were held by q disappear. Thus, after the move, the token $\mathbf{i}$ occurs at a state $p \in Q$ if and only if $p \in q.a$ for some state q that had held $\mathbf{i}$ just prior to the move. For an illustration, Fig. 1 (borrowed from [12]) demonstrates the initial case of a 5-state NFA with the input alphabet $\{0, 1\}$ (top), along with the outcomes of the first move, depending on whether 0 or 1 has been chosen for the move (bottom left and bottom right, respectively).

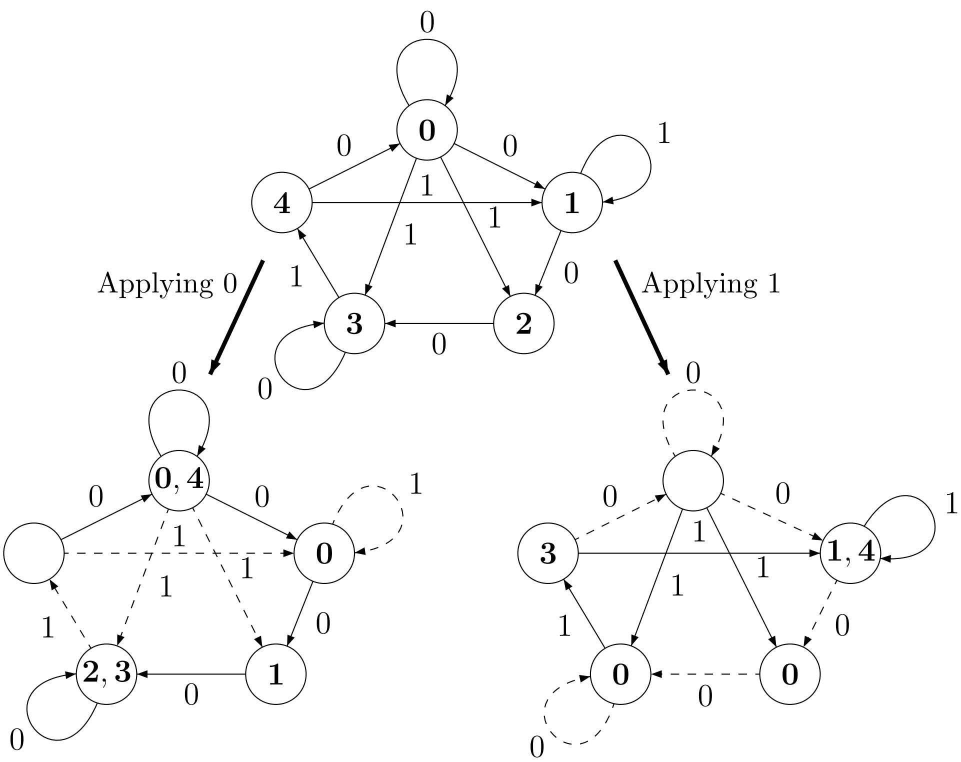


Figure 1. Redistribution of tokens after the first move

The intended meaning of the variables y_{ij}^t (which will be enforced by the condition we impose on them later) is as follows: $y_{ij}^t = 1$ should mean that after t rounds of the game Γ, one of the tokens held by the state q_j is $\mathbf{i}$.

Perhaps, it makes sense to add a matrix interpretation of the game Γ as the token variables get quite a clear meaning under this interpretation. The initial position of Γ can be thought of as the identity Boolean $Q \times Q$-matrix. At each move, an input symbol $a \in \Sigma$ is chosen and the matrix of the current position is right multiplied by the matrix $M(a)$. Then for each fixed t, the values of the variables y_{ij}^t are exactly the entries of the matrix corresponding to the position of Γ after t moves. For instance, the matrices that correspond to two possible positions of the game played on the 5-state NFA in Fig. 1 are

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Altogether, V consists of $n^2(\ell+1)+\ell$ variables so that we can take the polynomial $x^2(y+1)+y$ to play the role of $v(x,y)$ from the above definition of polynomial reduction. For the reduction from D3W to SAT in [12], an extra set of n variables (the so-called *synchronization variables*) was used. Here we have managed to slightly decrease the number of variables.

Now we describe the set C of clauses over V corresponding to the instance $(\mathscr{A}, \ell)$. As in [12], C is the disjoint union of set C_0 of *initial clauses*, the sets C_t, $t = 1, \dots, \ell$, of *transition clauses*, and the set S of *synchronization clauses*. The clauses in $C_0, C_1, \dots, C_\ell$ are constructed exactly as in [12] (but we will recall the construction for the reader's convenience) while the clauses in S are essentially different as these are the clauses that reflect the essence of D_2-synchronization.

The clauses in C_0 describe the initial position of the game Γ. In this position, each state $q_i \in Q$ holds the token $\mathbf{i}$ and nothing else. Therefore C_0 consists of n^2 one-literal clauses, namely, the clauses $y^0_{00}, \dots, y^0_{n-1\,n-1}$ along with all clauses of the form $\neg y^0_{ij}$ with $i \neq j$.

In order to define C_t for $t = 1, \dots, \ell$, consider for all $i, j = 0, \dots, n-1$, the following formulas:

$$\Psi^t_{ij}: \quad y^t_{ij} \iff \Big(x_t \wedge \bigvee_{q_k \in P_1(q_j)} y^{t-1}_{ik}\Big) \vee \Big(\neg x_t \wedge \bigvee_{q_h \in P_0(q_j)} y^{t-1}_{ih}\Big).$$

The equivalence Ψ^t_{ij} is nothing but a direct translation of the above propagation rule for the tokens in the language of propositional logic. Indeed, it says that the token $\mathbf{i}$ occurs at the state q_j after t moves if and only if one of the following alternatives takes place:

- the t-th move was done with the input symbol 1 and one of the preimages of q_j under the actions of 1 was holding $\mathbf{i}$ after $t-1$ moves, or
- the t-th move was done with the input symbol 0 and one of the preimages of q_j under the actions of 0 was holding $\mathbf{i}$ after $t-1$ moves.

The following fact is a special instance of [12, Lemma 2]:

Lemma 1. *Every truth assignment $\varphi\colon \{x_1, \dots, x_\ell\} \to \{0,1\}$ on the letter variables has a unique extension $\overline{\varphi}$ to the token variables y^t_{ij} that makes all the clauses in C_0 and all the formulas Ψ^t_{ij} hold true ($i, j = 0, \dots, n-1$, $t = 1, \dots, \ell$). The token variable y^t_{ij} gets value 1 under $\overline{\varphi}$ if and only if after the moves $\varphi(x_1), \dots, \varphi(x_t)$ of the game Γ, one of the tokens held by the state q_j is $\mathbf{i}$.*

Now, for each $t = 1, \dots, \ell$, we let C_t be the set of all clauses of a suitable CNF (conjunctive normal form) equivalent to $\bigwedge_{1 \le i,j \le n} \Psi^t_{ij}$. Of course, there are many ways to convert the latter formula to an equivalent CNF, but in order to reuse a part of code written for [12], we retain for C_t the following set of clauses:

$$\neg y^t_{ij} \vee x_t \vee \bigvee_{q_h \in P_0(q_j)} y^{t-1}_{ih}, \qquad \neg y^t_{ij} \vee \neg x_t \vee \bigvee_{q_k \in P_1(q_j)} y^{t-1}_{ik}, \tag{1.1}$$

$$y^t_{ij} \vee \neg x_t \vee \neg y^{t-1}_{ik} \quad \text{for each } q_k \in P_1(q_j), \tag{1.2}$$

$$y^t_{ij} \vee x_t \vee \neg y^{t-1}_{ih} \quad \text{for each } q_h \in P_0(q_j). \tag{1.3}$$

Clauses of the form (1.1)–(1.3) simplify if one of the sets $P_0(q_j)$ or $P_1(q_j)$ is empty. In (1.1) the disjunctions over the empty sets are omitted so that if, say, $P_0(q_j) = \varnothing$, then the first clause in (1.1) reduces to $\neg y^t_{ij} \vee x_t$. As for (1.2) or (1.3), these clauses disappear whenever $P_1(q_j)$ or, respectively $P_0(q_j)$ are empty. Thus, if the state q_j is such that $P_0(q_j) = P_1(q_j) = \varnothing$, then both (1.2) and (1.3) vanish and the two clauses in (1.1) reduce to $\neg y^t_{ij} \vee x_t$ and $\neg y^t_{ij} \vee \neg x_t$. The latter pair of clauses is clearly equivalent to just $\neg y^t_{ij}$ whence all clauses (1.1)–(1.3) reduce to $\neg y^t_{ij}$ for this particular j and for all $i = 0, \dots, n-1$ and $t = 1, \dots, \ell$. This fact amounts to expressing the following simple idea: if the state q_j has no incoming edges, then no token can arrive at q_j after any move of the game Γ.

Let m stand for the number of all *transitions in* $\mathscr{A}$, that is, triples $(q, a, q') \in Q \times \Sigma \times Q$ with $q' \in \delta(q, a)$. Clearly, $m \le 2n^2$. For each fixed i, the number $\sum_{j=1}^{n}(|P_1(q_j)| + |P_0(q_j)|)$ of clauses of the forms (1.2) and (1.3) is equal to m, whence the total number of such "short" clauses is mn. As for "long" clauses in (1.1), there are at most two such clauses for each fixed pair (i, j), whence their total number does not exceed $2n^2$. Altogether, $|C_t| \le n(m + 2n) \le 2n^2(n+1)$ for each $t = 1, \dots, \ell$.

While clauses in $\bigcup_{t=0}^{\ell} C_t$ coincide with those used in [12], the sets of synchronization clauses in [12] and here are different. The present set S contains n^2 disjunctions of the following form:

$$\neg y_{ij}^{\ell} \vee y_{i+1 \pmod{n}\, j}^{\ell}, \quad i,j = 0,\dots,n-1. \tag{1.4}$$

Clearly, for each fixed j, the clauses (1.4) are equivalent to the cycle of implications

$$y_{0j}^{\ell} \to y_{1j}^{\ell},\ y_{1j}^{\ell} \to y_{2j}^{\ell},\dots,y_{n-1\,j}^{\ell} \to y_{0j}^{\ell}$$

that expresses the idea of D_2-synchronization as follows: if the state q_j holds some token after ℓ moves, then q_j must hold all n tokens $\mathbf{0},\mathbf{1},\dots,\mathbf{n-1}$. Observe that the clauses (1.4) are satisfied if all variables y_{ij}^{ℓ} get value 0; by Lemma 1 this happens exactly when all tokens disappear after ℓ moves which means that the word $w \in \Sigma^{\ell}$ corresponding to the chosen sequence of moves is nowhere defined. In this paper we are interested in finding only those D_2-synchronizing words that are somewhere defined; we refer to them as *proper* D_2-synchronizing words. Therefore, we add to the set S the following clause:

$$\bigvee_{0\le j\le n-1} y_{0j}^{\ell}. \tag{1.5}$$

The clause (1.5) is satisfied if and only if some state holds the token $\mathbf{0}$ after ℓ moves; in the presence of (1.4), the latter fact is equivalent to the claim that some state holds *some* token after ℓ moves, which in turn means that the word w is somewhere defined.

The whole set $C = S \cup \bigcup_{t=0}^{\ell} C_t$ consists of at most $2n^2((n+1)\ell+1)+1$ clauses. Thus, the polynomial $2x^2((x+1)y+1)+1$ can be taken as $c(x,y)$ from the definition of polynomial reduction. Summarizing the above discussion, we arrive at the following result parallel to [12, Theorem 3].

Theorem 1. *An NFA $\mathscr{A}$ has a proper D_2-synchronizing word of length ℓ if and only if the instance (V,C) of SAT constructed above is satisfiable, and the construction takes time polynomial in the size of $\mathscr{A}$ and the value of ℓ. Moreover, a word $w = a_1\cdots a_\ell$ with $a_1,\dots,a_\ell \in \{0,1\}$ is proper D_2-synchronizing for $\mathscr{A}$ if and only if the map $x_t \mapsto a_t$, $t = 1,\dots,\ell$, extends to a satisfying assignment for (V,C).*

2. Experimental results

The general scheme of our experiments follows [12] mutatis mutandis. We outline our basic procedure, commenting on similarities with and differences from the procedure implemented in [12].

1. A positive integer n (the number of states) is fixed. As in [12], we have considered $n \le 100$.

2. A random NFA $\mathscr{A}$ with n states and 2 input symbols is generated. We have used the same two models of random generation that were used in [12] but we provide details below for the reader's convenience. As in [12], we disregard NFAs that have no everywhere defined input symbol because such NFAs possess neither D_3-synchronizing nor proper D_2-synchronizing words.

3. A positive integer ℓ_0 (the hypothetical length of the shortest D_2-synchronizing word for $\mathscr{A}$) is chosen. Taking into account the fact that proper D_2-synchronization is more restrictive than D_3-synchronization, we have used slightly larger values of ℓ_0 than in [12]. We introduce three integer variables $\ell_{\min}$, ℓ, and $\ell_{\max}$ and initialize them as follows: $\ell_{\min} := 1$, $\ell := \ell_0$, $\ell_{\max} := 2\ell_0$.

4. The pair $(\mathscr{A}, 1)$ is encoded into a SAT instance (V', C') as described in Sect. 1.

5. The instance (V', C') is scaled to the instance (V, C) that encodes the pair $(\mathscr{A}, \ell)$, see Remark 1 below.

6. The SAT solver MiniSat 2.2.0. is invoked to solve the SAT instance (V, C). We refer to [3] for a description of the underlying ideas of MiniSat and to [4] for a discussion and the source code of the solver.

7. The binary search on ℓ is performed. If MiniSat returns YES on the instance (V, C), we check whether or not $\ell = \ell_{\min}$. If $\ell = \ell_{\min}$, then ℓ is the minimum length of proper D_2-synchronizing words for $\mathscr{A}$, and we pass to Step 2 to generate another NFA. If $\ell > \ell_{\min}$, we keep the value of $\ell_{\min}$, update $\ell_{\max}$ and ℓ by letting

$$\ell_{\max} := \ell, \quad \ell := \left\lfloor \frac{\ell_{\min} + \ell_{\max}}{2} \right\rfloor,$$

   and pass to Step 5.

   If the MiniSat returns NO on the instance (V, C), we check whether or not $\ell = \ell_{\max}$. If $\ell = \ell_{\max}$, we interpret this as the evidence that the NFA $\mathscr{A}$ fails to be properly D_2-synchronizing[2] and go to Step 2 to generate another NFA. If $\ell < \ell_{\max}$, we keep the value of $\ell_{\max}$, update $\ell_{\min}$ and ℓ by letting

$$\ell_{\min} := \ell + 1, \quad \ell := \left\lceil \frac{\ell_{\min} + \ell_{\max}}{2} \right\rceil,$$

   and pass to Step 5.

Remark 1. In the course of the binary search outlined above, we have to consider instances of D2W with the same NFA $\mathscr{A}$ but different values of ℓ. An important feature of the encoding presented in Sect. 1 is that as soon as we have constructed the "primary" SAT instance (V', C') that encodes the D2W instance $(\mathscr{A}, 1)$, we are in a position to scale (V', C') to the SAT instance encoding the D2W instance $(\mathscr{A}, \ell)$ for any value of ℓ. In order to explain this feature, recall that MiniSAT accepts its input in the following text format (so-called simplified DIMACS CNF format). Every line beginning `c` is a comment. The first non-comment line is of the form:

```
p cnf NUMBER_OF_VARIABLES NUMBER_OF_CLAUSES
```

Variables are represented by integers from 1 to `NUMBER_OF_VARIABLES`. The first non-comment line is followed by `NUMBER_OF_CLAUSES` non-comment lines each of which defines a clause. Every such line starts with a space-separated list of different non-zero integers corresponding to the literals of the clause: a positive integer corresponds to a literal which is a variable, and a negative integer corresponds to a literal which is the negation of a variable; the line ends in a space and the number 0.

Given an NFA $\mathscr{A}$ with n states, we write the SAT instance (V', C'), which corresponds to $(\mathscr{A}, 1)$, in DIMACS CNF format, representing the variables y^0_{ij}, y^1_{ij}, and x_1 by the numbers, respectively, $in + j + 1$, $n^2 + in + j + 2$, and $n^2 + 1$. Consider, for a simple illustration, the NFA $\mathscr{E}_2$ shown in Fig. 2.

Table 1 in the next page presents our encoding of the D2W instance $(\mathscr{E}_2, 1)$ as a SAT instance. In the left column the SAT instance is shown as a list of clauses while the right column shows it in DIMACS CNF format.

[2] Of course, the equality $\ell = \ell_{\max}$ only means that $\mathscr{A}$ has no proper D_2-synchronizing word of length $\leq 2\ell_0$, and it is not excluded, in principle, that the NFA is properly D_2-synchronizing but its shortest proper D_2-synchronizing word is very long. However, by choosing appropriate values of the parameter ℓ_0, we have drastically minimized the number of the "bad" cases when the SAT solver returns NO and $\ell = \ell_{\max}$ so that we have been able to analyze each of them individually.

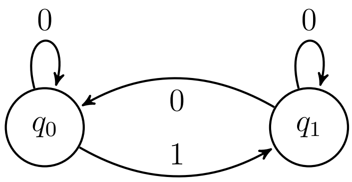


Figure 2. The NFA $\mathscr{E}_2$

	Clauses	**DIMACS CNF lines**
		`p cnf 9 25`
C'_0	y^0_{00}	`1 0`
	$\neg y^0_{01}$	`-2 0`
	$\neg y^0_{10}$	`-3 0`
	y^0_{11}	`4 0`
C'_1	$\neg y^1_{00} \vee x_1 \vee y^0_{00} \vee y^0_{01}$	`-6 5 1 2 0`
	$\neg y^1_{00} \vee \neg x_1$	`-6 -5 0`
	$y^1_{00} \vee x_1 \vee \neg y^0_{00}$	`6 5 -1 0`
	$y^1_{00} \vee x_1 \vee \neg y^0_{01}$	`6 5 -2 0`
	$\neg y^1_{01} \vee x_1 \vee y^0_{01}$	`-7 5 2 0`
	$\neg y^1_{01} \vee \neg x_1 \vee y^0_{00}$	`-7 -5 1 0`
	$y^1_{01} \vee \neg x_1 \neg y^0_{00}$	`7 -5 -1 0`
	$y^1_{01} \vee x_1 \vee \neg y^0_{01}$	`7 5 -2 0`
	$\neg y^1_{10} \vee x_1 \vee y^0_{10} \vee y^0_{11}$	`-8 5 3 4 0`
	$\neg y^1_{10} \vee \neg x_1$	`-8 -5 0`
	$y^1_{10} \vee x_1 \vee \neg y^0_{10}$	`8 5 -3 0`
	$y^1_{10} \vee x_1 \vee \neg y^0_{11}$	`8 5 -4 0`
	$\neg y^1_{11} \vee x_1 \vee y^0_{11}$	`-9 5 4 0`
	$\neg y^1_{11} \vee \neg x_1 \vee y^0_{10}$	`-9 -5 3 0`
	$y^1_{11} \vee \neg x_1 \neg y^0_{10}$	`9 -5 -3 0`
	$y^1_{11} \vee x_1 \vee \neg y^0_{11}$	`9 5 -4 0`
S'	$\neg y^1_{00} \vee y^1_{01}$	`-6 7 0`
	$\neg y^1_{01} \vee y^1_{00}$	`-7 6 0`
	$\neg y^1_{10} \vee y^1_{11}$	`-8 9 0`
	$\neg y^1_{11} \vee y^1_{10}$	`-9 8 0`
	$y^1_{00} \vee y^1_{01}$	`6 7 0`

Table 1. The SAT encoding of the D2W instance $(\mathscr{E}_2, 1)$

Now, in order to scale (V', C') to the SAT instance (V, C) that encodes the pair $(\mathscr{A}, \ell)$ for some given $\ell > 1$, we perform the following transformations on the DIMACS CNF representation of $C' = C'_0 \cup C'_1 \cup S'$:

1. In the first non-comment line, replace `NUMBER_OF_VARIABLES` and `NUMBER_OF_CLAUSES` by

$n^2(\ell+1)+\ell$ and respectively $\ell N + 2n^2 + 1$, where N is the number of clauses in C_1'.

2. Keep the lines corresponding to the clauses in C_0' and C_1'.

3. For each $t = 2, \ldots, \ell$, add all the lines obtained from the ones corresponding to the clauses C_1' by keeping the sign of every non-zero integer and adding $(t-1)n^2 + t - 1$ to its absolute value.

4. In each line corresponding to a clause in S', substitute every nonzero integer $\pm k$ by the integer $\pm(k + (\ell-1)n^2 + \ell - 1)$.

Our experiments have been performed on a personal computer equipped with an Intel(R) Core(TM) i5-2520M processor with 2.5 GHz CPU and 4GB of RAM. We have implemented the described algorithm in C++ and compiled with GCC 4.9.2. For various fixed $n \leq 100$, up to 1000 NFAs with n states have been generated and analyzed. We have generated 1000 automata for each $n \in \{5, 10, \ldots, 30\}$, 700 automata for each $n \in \{35, 40, \ldots, 60\}$, 500 automata for each $n \in \{65, 70, \ldots, 80\}$, and 200 automata for each $n \in \{90, 100\}$. The calculations have taken ≈ 400 seconds for $n = 10$ and $\approx 1.2 \cdot 10^5$ seconds for $n = 80$.

As in [12], the two models used for random generation of an NFA $\mathscr{A} = (Q, \Sigma, \delta)$ with n states and 2 input symbols were the *uniform model* based on the uniform distribution and the *Poisson model* based on the Poisson distribution with some parameter λ. For each state $q \in Q$ and each symbol $s \in \Sigma$, we first choose a number $k \in \{0, 1, 2, \ldots, n\}$ that serves as the cardinality of the set $\delta(q, s)$. In the uniform model, each k is chosen with probability $1/(n+1)$ while in the Poisson model with parameter λ, each $k < n$ is chosen with probability $e^{-\lambda}\lambda^k/k!$ and n is chosen with probability $1 - e^{-\lambda}\sum_{k=0}^{n-1}\lambda^k/k!$. With k having been chosen, we proceeded the same in both models, by choosing $\delta(q, s)$ from all $\binom{n}{k}$ subsets of Q with cardinality k uniformly at random.

For NFAs generated under the uniform model, we have observed that for an overwhelming majority of properly D_2-synchronizing NFAs, the length of the shortest proper D_2-synchronizing word is 3, and this conclusion does not depend on the number of states within the range of our experiments. Recall that the experiments in [12] revealed quite a similar phenomenon for D_3-synchronization: if a NFA generated under the uniform model is D_3-synchronizing (which happens with the probability $\approx 60\%$, see [12, Proposition 5]), then its shortest D_3-synchronizing word has length 2, and this fact does not depend on the number of states. An informal explanation of the latter phenomenon can be found in [12]; similar arguments apply also in the present situation.

Thus, the uniform model fails to produce any "slowly synchronizing" NFA. This indicates that using SAT-solvers in the uniform setting was not really necessary since a brute-force approach would suffice. Indeed, given an NFA $\mathscr{A} = (Q, \Sigma, \delta)$, one can write all words over Σ up to a given length in the short-lex order and apply each of these words to $\mathscr{A}$ until one finds a D_2-synchronizing word. As our experiments reveal, for a majority of NFAs generated under the uniform model, the brute-force approach requires to check only words up to length 3.

For random NFAs generated under the Poisson model, our experiments show that if the parameter λ is fixed, the length of the shortest proper D_2-synchronizing word grows with the number of states but the growth rate is rather small. Some sample experimental results are presented in Fig. 3. The three graphs in Fig. 3 correspond to NFAs with 30, 45, and 60 states generated under the Poisson models with $\lambda = 2$ and demonstrate how these NFAs are distributed according to the length of their shortest proper D_2-synchronizing words. The horizontal axis is the minimum length of proper D_2-synchronizing words and the vertical axis is the number of NFAs. We have applied the method of least squares to our experimental data, searching for an explicit function of n that approximates the mean value $E_\lambda(n)$ of the minimum lengths of proper D_2-synchronizing words for n-state NFAs generated under the Poisson model with a given parameter λ. The best

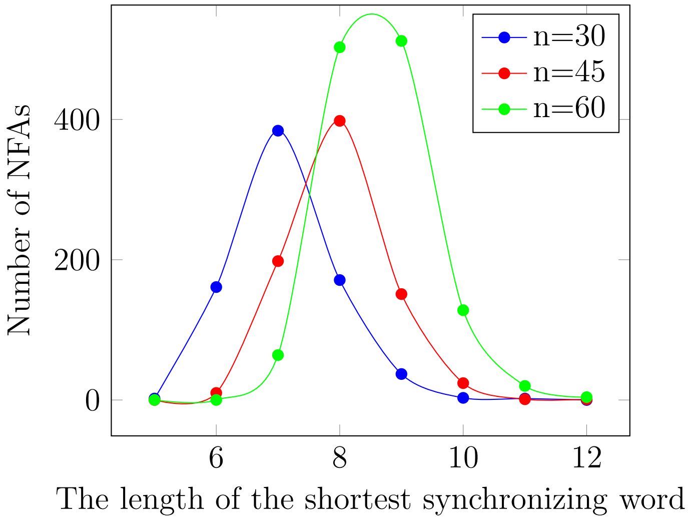


Figure 3. Distributions of random NFAs with 30, 45, and 60 states generated under the Poisson model with $\lambda = 2$ according to the minimum lengths of their proper D_2-synchronizing words

approximations have been provided by logarithmic functions; for instance, for $\lambda = 2$, we have found the following solution:

$$E_2(n) \approx -0.39 + 2.2\ln(n).$$

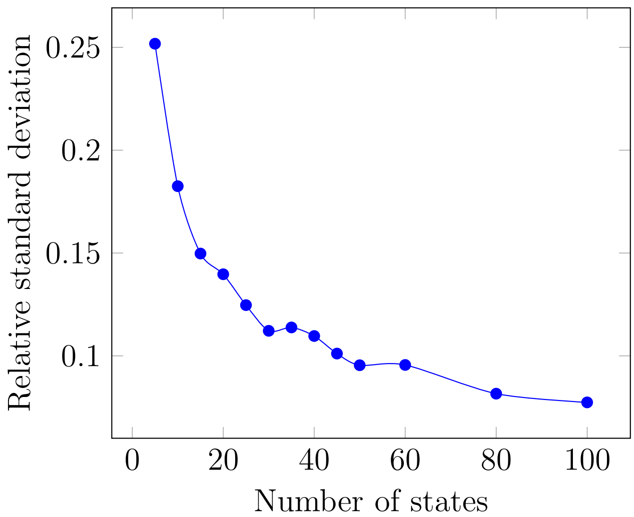


Figure 4. The relative standard deviation of the minimum lengths of proper D_2-synchronizing words for n-state NFAs as a function of n

Fig. 4 shows the relation between the relative standard deviation of our datasets and the number of states (for $\lambda = 2$).

Besides experimenting with randomly generated NFAs, we have tested our approach on certain provably "slowly synchronizing" NFAs considered in the literature. Here we report a set of results in which we used as benchmarks several automata from the series $\mathscr{P}_n$ suggested by de Bondt, Don, and Zantema [2]. The state set of $\mathscr{P}_n$ is $\{1, 2, \ldots, n\}$, $n \geq 3$, and the input alphabet consists of two letters a and b whose actions are defined as follows:

$$q.a := \begin{cases} q+1 & \text{if } q = 1,2, \\ q & \text{if } q = 3,\dots,n; \end{cases} \qquad q.b := \begin{cases} \text{undefined} & \text{if } q = 1, \\ q+1 & \text{if } q = 2,\dots,n-1, \\ 1 & \text{if } q = n. \end{cases}$$

Thus, the automata $\mathscr{P}_n$ are partial deterministic, and it is easy to see that for partial deterministic automata, proper D_2-synchronizing words coincide with D_3-synchronizing words and coincide with so-called carefully synchronizing words considered in [2]. Hence we can compare the information about the length of shortest synchronizing words for $\mathscr{P}_n$ obtained in [2, Theorem 3] and the results produced by an application of our procedure. In our experiments, we have examined all automata $\mathscr{P}_n$ with $n = 4, 5, \dots, 11$, and for each of them, our result has matched the theoretical value predicted by [2, Theorem 3]. The time consumed ranges from 0.301 sec for $n = 4$ to 4303 sec for $n = 11$. Observe that in the latter case the shortest synchronizing word has length 116; clearly, this value is out of reach for any brute-force method.

3. Conclusion and future work

We have presented a modification of the approach originated in [12] that has allowed us to find shortest proper D_2-synchronizing words for nondeterministic automata with two input letters and up to 100 states. The size of automata that we are able to analyze may seem modest in comparison with the results of [8] whose authors describe sophisticated methods to compute shortest synchronizing words for *deterministic* automata with up to 350 states. However, two important nuances should be taken into account. First, for the time being, the approach of [12] and the present paper appears to be the only one that has proved to work in the realm of nondeterministic automata. Second, it is well known that nondeterministic automata may be exponentially more succinct than equivalent deterministic ones, and, say, an NFA with 100 states may encode the same amount of information as a DFA with 2^{100} states.

We have concentrated on D_2-synchronizing words which are everywhere defined. In fact, shortest nowhere defined words are even easier to be find with a similar method. The point is that in terms of our game Γ from Sect. 1, a nowhere defined word is just a word which application removes all tokens. However, if all tokens are going to be eventually removed, there is no need to distinguish between them! Therefore one can drastically reduce the number of variables and clauses used in the encoding. Instead of the 3-parameter set of variables $\{y_{ij}^t\}$ used in Sect. 1, it suffices to consider the 2-parameter set $\{y_j^t\}$ where $y_j^t = 1$ should mean that after t rounds of the game Γ, the state q_j holds a token; similarly, the role of the 3-parameter set of formulas $\{\Psi_{ij}^t\}$ can be played by the 2-parameter set consisting of the formulas

$$y_j^t \iff \Big(x_t \wedge \bigvee_{q_k \in P_1(q_j)} y_k^{t-1}\Big) \vee \Big(\neg x_t \wedge \bigvee_{q_h \in P_0(q_j)} y_h^{t-1}\Big)$$

for all $j = 0, \dots, n-1$ and $t = 1, \dots, \ell$. Similar simplifications apply to the sets of initial and synchronization clauses. Therefore, it made no sense to search for nowhere and everywhere defined D_2-synchronizing words simultaneously, although it was possible (for this, one just had to remove the clause (1.5) from the set of synchronization clauses).

Yet another version of synchronization for nondeterministic automata suggested in [7] is D_1-synchronization. A word $w \in \Sigma^*$ is said to be D_1-*synchronizing* for $\mathscr{A} = (Q, \Sigma, \delta)$ if $q.w = q'.w$ and $|q.w| = 1$ for all $q, q' \in Q$. Clearly, every D_1-synchronizing word is everywhere defined and is D_2-synchronizing but the converse is not true: an everywhere defined D_2-synchronizing word need not be D_1-synchronizing. We can use encodings similar to those in [12] and the present paper in order to find shortest D_1-synchronizing words for NFAs of reasonable sizes; one only has to adjust the set of synchronization clauses.

We think that the results presented here and in [12] demonstrate that our approach works in principle but, of course, its present implementation is only a toy prototype for a system that could be used for real-world applications. There are several resources, on both software and hardware sides, which can be employed to speed up our calculations and enlarge their range. In particular, one can try more advanced SAT-solvers, such as CryptoMiniSat [14] and lingeling [1], and run a version of our program on a multiprocessor grid.

Acknowledgments. The author thanks the anonymous referees for their constructive comments and recommendations.

REFERENCES

1. Biere A. Yet another local search solver and lingeling and friends entering the SAT Competition 2014. In: *Proceedings of SAT Competition 2014: Solver and Benchmark Descriptions*. University of Helsinki, 2014. P. 39–40. URL: http://fmv.jku.at/papers/Biere-SAT-Competition-2014.pdf
2. de Bondt M., Don H., Zantema H. *Lower bounds for synchronizing word lengths in partial automata.* Preprint, 2018. URL: https://arxiv.org/abs/1801.10436
3. Eén N., Sörensson N. An extensible SAT-solver. *Lect. Notes Comput. Sci.*, Vol. 2919: Theory and Applications of Satisfiability Testing (SAT 2003). 2004. P. 502–518. DOI: 10.1007/978-3-540-24605-3_37
4. Eén N., Sörensson N. *The MiniSat Page.* URL: http://minisat.se
5. Gomes C. P., Kautz H., Sabharwal A., Selman B. Satisfiability Solvers. Ch. 2. In: *Handbook of Knowledge Representation*, Elsevier, 2008. P. 89–134. DOI: 10.1016/S1574-6526(07)03002-7
6. Güniçen C., Erdem E., Yenigün H. Generating shortest synchronizing sequences using Answer Set Programming. In: *Proceedings of Answer Set Programming and Other Computing Paradigms (ASPOCP 2013).* P. 117–127. URL: https://arxiv.org/abs/1312.6146.
7. Imreh B., Steinby M. Directable nondeterministic automata. *Acta Cybernetica*. 1999. Vol. 14, no. 1. P. 105–115.
8. Kisielewicz A., Kowalski J., Szykuła M. Computing the shortest reset words of synchronizing automata. *J. Comb. Optim.* 2015. Vol. 29, no. 1. P. 88–124. DOI: 10.1007/s10878-013-9682-0
9. Martyugin P. Synchronization of automata with one undefined or ambiguous transition. *Lect. Notes Comput. Sci.*, Vol. 7381: Implementation and Application of Automata (CIAA 2012). 2012. P. 278–288. DOI: 10.1007/978-3-642-31606-7_24
10. Rystsov I. K. Polynomial complete problems in automata theory. *Inf. Process. Lett.* 1983. Vol. 16, no. 3. P. 147–151. DOI: 10.1016/0020-0190(83)90067-4
11. Rystsov I. K. Asymptotic estimate of the length of a diagnostic word for a finite automaton. *Cybernetics*. 1980. Vol. 16, no. 1. P. 194–198. DOI: 10.1007/bf01069104
12. Shabana H., Volkov M. V. Using Sat solvers for synchronization issues in nondeterministic automata. *Siberian Electronic Math. Reports*. 2018. Vol. 15. P. 1426–1442. URL: http://semr.math.nsc.ru/v15/p1426-1442.pdf
13. Skvortsov E., Tipikin E. Experimental study of the shortest reset word of random automata. *Lect. Notes Comput. Sci.*, Vol. 6807: Implementation and Application of Automata (CIAA 2011), 2011. P. 290–298. DOI: 10.1007/978-3-642-22256-6_27
14. Soos M. CryptoMiniSat 2. URL: http://www.msoos.org/cryptominisat2/

Amendments to my article
THE EKATERINBURG SEMINAR "ALGEBRAIC SYSTEMS": 50 YEARS OF ACTIVITIES

Published in: Ural Mathematical Journal. Vol. 3, No. 1, 2017, pp. 3–26
DOI: 10.15826/umj.2017.1.008

In the article indicated, the following two omissions and a misprint need to be mended. They originally appeared in the preceding version of a jubilee article about the seminar published in 2007 (see the reference [5] in the article under revision) and have been reproduced by the author in the latter version because of an oversight.

Namely, on p. 7, item 10, and on p. 8, item 18, the co-supervisors L. G. Mustafaev and, respectively, R. A. Bairamov should be mentioned as well; on p. 25, line 3, there must be "works" instead of "words".

Lev N. Shevrin
lev.shevrin@urfu.ru

Editor: Tatiana F. Filippova

Managing Editor: Oxana G. Matviychuk

Design: Alexander R. Matvichuk

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990

Phone: +7 (343) 375-34-73

Fax: +7 (343) 374-25-81

Email: secretary@umjuran.ru

Web-site: https://umjuran.ru

N.N.Krasovskii Institute of Mathematics and Mechanics
of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N.Yeltsin

Distributed for free