БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ЖУРНАЛ
БЕЛОРУССКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

МАТЕМАТИКА ИНФОРМАТИКА

JOURNAL
OF THE BELARUSIAN STATE UNIVERSITY

MATHEMATICS and INFORMATICS

Издается с января 1969 г.
(до 2017 г. – под названием «Вестник БГУ.
Серия 1, Физика. Математика. Информатика»)

Выходит три раза в год

2

2021

МИНСК
БГУ

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

EDITORIAL BOARD

Вещественный, комплексный и функциональный анализ

Real, complex and functional analysis

УДК 517.938

УПАКОВОЧНЫЕ РАЗМЕРНОСТИ БАССЕЙНОВ, ПОРОЖДЕННЫХ РАСПРЕДЕЛЕНИЯМИ НА КОНЕЧНОМ АЛФАВИТЕ

В. И. БАХТИН[1)]***, Б. САДОК***[2)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Люблинский католический университет им. Иоанна Павла II, ал. Рацлавицкая, 14, 20-950, г. Люблин, Польша*

Рассматривается пространство бесконечных сигналов, составленных из букв конечного алфавита. Каждый сигнал порождает последовательность эмпирических мер на алфавите и отвечающее этой последовательности предельное множество. Все пространство сигналов разбивается на узкие бассейны, состоящие из сигналов с одинаковыми предельными множествами для последовательности эмпирических мер. Для каждого узкого бассейна вычисляется его упаковочная размерность. Кроме того, рассчитываются упаковочные размерности бассейнов двух других типов, определяемых в терминах предельного поведения эмпирических мер.

Ключевые слова: упаковочная размерность; эмпирическая мера; бассейн вероятностной меры.

Образец цитирования:
Бахтин ВИ, Садок Б. Упаковочные размерности бассейнов, порожденных распределениями на конечном алфавите. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:6–16 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-6-16

For citation:
Bakhtin VI, Sadok B. Packing dimensions of basins generated by distributions on a finite alphabet. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:6–16.
https://doi.org/10.33581/2520-6508-2021-2-6-16

Авторы:
Виктор Иванович Бахтин – доктор физико-математических наук, профессор; профессор кафедры функционального анализа и аналитической экономики механико-математического факультета.
Бруно Садок – лектор кафедры теории вероятностей и статистики факультета естественных наук и наук о здоровье.

Authors:
Victor I. Bakhtin, doctor of science (physics and mathematics), full professor; professor at the department of functional analysis and analytical economics, faculty of mechanics and mathematics.
bakhtin@tut.by
https://orcid.org/0000-0001-5782-5378
Bruno Sadok, lecturer at the department of probability theory and statistics, faculty of natural sciences and health.
bruno.bonitas@gmail.com

PACKING DIMENSIONS OF BASINS GENERATED BY DISTRIBUTIONS ON A FINITE ALPHABET

V. I. BAKHTIN[a], *B. SADOK*[b]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[b]*John Paul II Catholic University of Lublin, 14 Racławickie Alley, Lublin 20-950, Poland*
Corresponding author: V. I. Bakhtin (bakhtin@tut.by)

We consider a space of infinite signals composed of letters from a finite alphabet. Each signal generates a sequence of empirical measures on the alphabet and the limit set corresponding to this sequence. The space of signals is partitioned into narrow basins consisting of signals with identical limit sets for the sequence of empirical measures and for each narrow basin its packing dimension is computed. Furthermore, we compute packing dimensions for two other types of basins defined in terms of limit behaviour of the empirical measures.

Keywords: packing dimension; empirical measure; basin of a probability measure.

Introduction

Signals of infinite length composed of letters from a finite alphabet may be classified in accordance with limit behaviour of generated by these signals empirical measures on the alphabet. It turns out that different classes of signals (which we call basins) have a sophisticated fractal structure, and the most adequate quantitative characteristic for their description is fractal dimension. For the first time Hausdorff dimensions of certain basins were calculated by Billingsley in [1; 2]. In [3] it was suggested to consider the so-called narrow basins, that are distinguished among other types of basins by the fact that they form a partition of the space of all infinite signals, and their Hausdorff dimensions were calculated. After Hausdorff dimension the next mostly used fractal dimension is the packing one. In [4] the authors announced explicit formulae for the packing dimensions of narrow basins and basins of certain other types. In the present paper we set forth detailed proofs for those formulae.

Let us proceed to strict definitions and statements.

Consider a finite set $X = \{1, \ldots, k\}$. In what follows this set will be called the *alphabet* and its elements the *letters*. Any sequences of letters, finite or infinite, will be called the *signals*. The set of finite signals of length n is naturally denoted as X^n, and the set of all infinite signals as

$$X^{\mathbb{N}} = \{x = (x_1, x_2, \ldots) | x_i \in X\}.$$

Each initial segment of a signal will be called its prefix.

Let $M(X)$ be the set of all probability measures on X:

$$M(X) = \left\{\mu = (\mu(1), \ldots, \mu(k)) \in \mathbb{R}^k \,\middle|\, \sum_i \mu(i) = 1, \mu(i) \geq 0\right\}.$$

Evidently, $M(X)$ is convex and compact. For each letter $x \in X$ denote by δ_x the unit measure supported at x, that is

$$\delta_x(y) = \begin{cases} 1, \text{ if } y = x, \\ 0, \text{ if } y \neq x. \end{cases}$$

Every finite signal $x = (x_1, \ldots, x_n)$ generates an *empirical measure* $\delta_{x,n} \in M(X)$ by the rule

$$\delta_{x,n} = \frac{\delta_{x_1} + \ldots + \delta_{x_n}}{n}.$$

In other words, $\delta_{x,n}(y)$ is the average frequency of the letter y among $x_1, \ldots, x_n$. Every infinite signal $x = (x_1, x_2, \ldots) \in X^{\mathbb{N}}$ defines a sequence of empirical measures $\delta_{x,n}$ generated by its prefixes of length n.

For each infinite signal x denote by $V(x)$ the set of all limit points of the sequence $\delta_{x,n} \in M(X)$. In view of compactness of $M(X)$ this set is non-empty. Moreover, in [3, lemma 3] it is proved that $V(x)$ is compact and connected.

For every subset $W \subset M(X)$ let us define the following sets in $X^{\mathbb{N}}$: the basin $\mathrm{B}(W)$, narrow basin $\mathrm{NB}(W)$, and wide basin $\mathrm{WB}(W)$ by formulae

$$\mathrm{B}(W) = \left\{x \in X^{\mathbb{N}} \,\middle|\, V(x) \subset W\right\},$$

$$\mathrm{NB}(W) = \left\{x \in X^{\mathbb{N}} \,\middle|\, V(x) = W\right\},$$

$$\mathrm{WB}(W) = \left\{x \in X^{\mathbb{N}} \,\middle|\, V(x) \cap W \neq \varnothing\right\}.$$

In other words, $\mathrm{B}(W)$ denotes the set of infinite signals x such that all limit points of the sequence of empirical measures $\delta_{x,n}$ belong to W, $\mathrm{NB}(W)$ denotes the collection of infinite signals x such that the set of limit points of the sequence $\delta_{x,n}$ coincides with W, and $\mathrm{WB}(W)$ denotes the set of infinite signals x such that the sequence $\delta_{x,n}$ has at least one limit point in W. Obviously, these basins satisfy the inclusions

$$\mathrm{NB}(W) \subset \mathrm{B}(W) \subset \mathrm{WB}(W).$$

From the above mentioned compactness and connectedness of $V(x)$ it follows that a narrow basin $\mathrm{NB}(W)$ may be non-empty only in the case when the corresponding set of limit points W is non-empty, compact, and connected. Conversely, in [3] it was proved that for every non-empty connected compact set $W \subset M(X)$ the narrow basin $\mathrm{NB}(W)$ is indeed non-empty. As for basins $\mathrm{B}(W)$ and $\mathrm{WB}(W)$, it is easily seen that they are non-empty for all $W \neq \varnothing$.

Every infinite signal x defines uniquely the set $V(x)$. Therefore the narrow basins $\mathrm{NB}(W)$ corresponding to different limit sets W do not intersect each other. Thus the entire space of infinite signals turns out to be partitioned into the narrow basins corresponding to different connected compact subsets $W \subset M(X)$. However, the basins of two other types may have non-empty intersections.

Let us fix a row of numbers $\theta = (\theta(1), \theta(2), \ldots, \theta(k)) \in (0, 1)^k$ (one number $\theta(i)$ for each letter $i \in X$) and define a metric ρ on the space of infinite signals $X^{\mathbb{N}}$ in the following way:

$$\rho(x, y) = \prod_{t=1}^{n} \theta(x_t), \text{ where } n = \inf\left\{t \,|\, x_t \neq y_t\right\} - 1. \tag{1}$$

Here n denotes the length of the largest common prefix of x and y. If $n = 0$ then we put $\rho(x, y) = 1$.

Consider the function

$$S(\mu, \theta) = \frac{\sum_{i=1}^{k} \mu(i) \ln \mu(i)}{\sum_{i=1}^{k} \mu(i) \ln \theta(i)}, \ \mu \in M(X).$$

It is easy to see that it depends continuously on μ (under the convention $0 \ln 0 = 0$).

The purpose of this paper is to prove the following two theorems declared in [4].

Theorem 1. *Suppose the space $X^{\mathbb{N}}$ is equipped with metric (1). Then for any non-empty connected compact subset $W \subset M(X)$ we have the equality*

$$\dim_P \mathrm{NB}(W) = \sup_{\mu \in W} S(\mu, \theta),$$

where $\dim_P$ *denotes the packing dimension.*

Theorem 2. *For any non-empty subset $W \subset M(X)$ we have*

$$\dim_P \mathrm{B}(W) = \sup_{\mu \in W} S(\mu, \theta), \tag{2}$$

$$\dim_P \mathrm{WB}(W) = \dim_P X^{\mathbb{N}} = \sup_{\mu \in M(X)} S(\mu, \theta). \tag{3}$$

Remark. Hausdorff dimensions of the basins $\mathrm{B}(W)$ were first calculated in [1; 2] (under the additional assumption $\sum_i \theta(i) = 1$, which may in fact be omitted), and dimensions of the wide basins were calculated in [5]. They have the form

$$\dim_H \mathrm{WB}(W) = \dim_H \mathrm{B}(W) = \sup_{\mu \in W} S(\mu, \theta).$$

Hausdorff dimensions of the narrow basins were recently calculated in [3]:

$$\dim_H \mathrm{NB}(W) = \inf_{\mu \in W} S(\mu, \theta).$$

Concerning the packing dimensions of basins, as far as we know, they were not investigated by anyone earlier.

The paper has the following structure. In the section «Packing dimensions of sets and local dimensions of measures» we define the packing dimensions of sets, local dimensions of measures, and formulate a theorem about relationships between them. In the section «An upper estimate for the packing dimension of a basin» we prove an upper estimate for the packing dimension of a basin. In the section «Construction of a model set of signals» we construct a model set of signals contained in the narrow basin. In the last section «A lower estimate for the packing dimension of a narrow basin», using the local dimensions of measures, we prove a lower estimate for the packing dimension of the model set and then deduce from it theorems 1 and 2.

Packing dimensions of sets and local dimensions of measures

At first we recall definitions of packing measures and dimensions. A packing of a set A in a metric space is any finite or countable collection of balls $B(x_i, r_i)$, centered at $x_i \in A$ and of radii r_i, such that $\rho(x_i, x_j) > r_i + r_j$ for all $i \neq j$. An ε-packing is a packing consisting of balls with radii not greater than ε.

For every $s > 0$ we put

$$C_\varepsilon^s(A) = \sup\left\{ \sum_i r_i^s \,\middle|\, \text{balls } B(x_i, r_i) \text{ form an } \varepsilon\text{-packing of } A \right\}.$$

Evidently, $C_\varepsilon^s(A)$ does not increase while ε decreases, and therefore there exists a limit

$$C^s(A) = \lim_{\varepsilon \to 0} C_\varepsilon^s(A).$$

The packing measure of dimension s of a set A is

$$P^s(A) = \inf\left\{ \sum_i C^s(A_i) \,\middle|\, \text{the sets } A_i \text{ form a countable cover of } A \right\},$$

and its packing dimension is defined as

$$\dim_P A = \inf\left\{ s > 0 \,\middle|\, P^s(A) = 0 \right\}.$$

Let M be a metric space and μ be a Borel measure on M. Then the function

$$D_\mu(x) = \limsup_{r \to 0+0} \frac{\ln \mu(B(x, r))}{\ln r}, \quad x \in M,$$

is called the upper local dimension of the measure μ.

The next theorem enables to calculate the packing dimensions of sets by means of the local dimensions of measures.

Theorem 3 [6, proposition 2.3]. *Suppose A is a subset of a metric space M. If there exists a finite Borel measure μ on M such that $D_\mu(x) \le s$ for all $x \in A$, then $\dim_P A \le s$. Conversely, if $D_\mu(x) \ge s$ for all $x \in A$ and the outer measure $\mu^*(A)$ is positive, then $\dim_P A \ge s$.*

An upper estimate for the packing dimension of a basin

Now for any non-empty subset $W \subset M(X)$ we prove the estimate

$$\dim_P \mathrm{B}(W) \le \sup_{\mu \in W} S(\mu, \theta). \tag{4}$$

It will imply the same estimate for the packing dimension of the narrow basin $\mathrm{NB}(W)$, since the latter is contained in $\mathrm{B}(W)$.

For each infinite signal $x = (x_1, x_2, \ldots)$ and positive integer n we define a cylinder $Z_n(x)$ as the set of all infinite signals with prefix $(x_1, \ldots, x_n)$:

$$Z_n(x) = \left\{ y = (y_1, y_2, \ldots) \in X^{\mathbb{N}} \,\middle|\, y_1 = x_1, \ldots, y_n = x_n \right\}.$$

Denote by $|Z_n(x)|$ its diameter with respect to metric (1). Obviously, it can be computed according to the formula

$$|Z_n(x)| = \prod_{t=1}^{n} \theta(x_t),$$

and its logarithm can be written in the form

$$\ln|Z_n(x)| = \prod_{t=1}^{n} \ln\theta(x_t) = n\sum_{i=1}^{k} \delta_{x,n}(i)\ln\theta(i). \tag{5}$$

Recall that $M(X)$ consists of all probability measures on the alphabet $X = \{1, \ldots, k\}$. These measures can be interpreted as vectors $\mu = (\mu(1), \ldots, \mu(k))$ in $\mathbb{R}^k$. Supply the space $\mathbb{R}^k$ with the norm

$$\|\mu\| = \sum_{i=1}^{k} |\mu_i|, \text{ where } \mu = (\mu_1, \ldots, \mu_k).$$

It naturally defines a metric and topology on $M(X)$.

For any neighbourhood $O(\mu)$ of a measure $\mu \in M(X)$ let us define the sets

$$X^n(O(\mu)) = \left\{x = (x_1, \ldots, x_n) \in X^n \,\middle|\, \delta_{x,n} \in O(\mu)\right\}, \ n \in \mathbb{N}. \tag{6}$$

By McMillan's equipartition theorem [7, p. 51] for any measure $\mu \in M(X)$ and any $\varepsilon > 0$ there exists a neighbourhood $O(\mu)$ and a number $N(\mu, \varepsilon)$ such that

$$\operatorname{card} X^n(O(\mu)) \le e^{n(h(\mu)+\varepsilon)} \text{ for all } n \ge N(\mu, \varepsilon), \tag{7}$$

where $h(\mu)$ is the entropy of μ defined as

$$h(\mu) = -\sum_{i=1}^{k} \mu(i)\ln\mu(i).$$

Now we start to prove inequality (4). Set

$$c_\theta = \min_{1 \le i \le k} |\ln\theta(i)|. \tag{8}$$

Fix an arbitrary number s satisfying the condition

$$s > \sup_{\mu \in W} S(\mu, \theta),$$

and choose $\varepsilon > 0$ so small that

$$\sup_{\mu \in W} S(\mu, \theta) < s - \frac{3\varepsilon}{c_\theta}. \tag{9}$$

Henceforth we will consider measures μ belonging to the closure $\bar{W}$ of the set W. For each $\mu \in \bar{W}$ choose a neighbourhood $O(\mu)$ sufficiently small to satisfy condition (7) and, in addition, such that for all measures $\nu \in O(\mu)$ the following inequality holds

$$\frac{\sum_{i=1}^{k} \mu(i)\ln\mu(i)}{\sum_{i=1}^{k} \nu(i)\ln\theta(i)} < \frac{\sum_{i=1}^{k} \mu(i)\ln\mu(i)}{\sum_{i=1}^{k} \mu(i)\ln\theta(i)} + \frac{\varepsilon}{c_\theta} = S(\mu, \theta) + \frac{\varepsilon}{c_\theta}. \tag{10}$$

Then from (8)–(10) it follows that

$$\frac{\sum_{i=1}^{k} \mu(i)\ln\mu(i)}{\sum_{i=1}^{k} \nu(i)\ln\theta(i)} < s - \frac{2\varepsilon}{c_\theta} \le s + \frac{2\varepsilon}{\sum_{i=1}^{k} \nu(i)\ln\theta(i)},$$

wherefrom, after multiplication by the negative denominator, we obtain

$$s\sum_{i=1}^{k}\nu(i)\ln\theta(i) < -h(\mu) - 2\varepsilon \text{ for all } \nu \in O(\mu). \tag{11}$$

Notice that if for some $x \in X^{\mathbb{N}}$ we have $\delta_{x,n} \in O(\mu)$ then by (5) and (11)

$$s\ln|Z_n(x)| = sn\sum_{i=1}^{k}\delta_{x,n}(i)\ln\theta(i) < -n\big(h(\mu) + 2\varepsilon\big),$$

and hence

$$|Z_n(x)|^s < e^{-n\left(h(\mu)+2\varepsilon\right)}. \tag{12}$$

Thus, for every measure $\mu \in \bar{W}$ there exists a neighbourhood $O(\mu)$ such that conditions (7), (12) hold simultaneously. Choose a finite cover $O(\mu_1)$, …, $O(\mu_l)$ of the compact set $\bar{W}$ by neighbourhoods of that type.

Consider a sequence of sets

$$A_N = \big\{x \in \mathrm{B}(W) \big| \delta_{x,n} \in O(\mu_1) \cup \ldots \cup O(\mu_l) \text{ for all } n \ge N\big\}. \tag{13}$$

Evidently, the greater is N, the greater is A_N. The definition of a basin implies that for each signal $x \in \mathrm{B}(W)$ the distance from $\delta_{x,n}$ to W tends to zero when $n \to \infty$. It follows that the sets A_N form a cover of the basin $\mathrm{B}(W)$.

Take any positive integers m, N satisfying the conditions

$$m \ge N \ge \max_{1 \le j \le l} N(\mu_j, \varepsilon),$$

where $N(\mu_j, \varepsilon)$ are the constants from (7) corresponding to the measures μ_j. Consider an arbitrary packing of the set A_N by disjoint cylinders of the form $Z_{n_i}(x_i)$, where $x_i \in A_N$ and $n_i \ge m$. For each $n \ge N(\mu_j, \varepsilon)$ the number of different cylinders $Z_n(x)$ such that $\delta_{x,n} \in O(\mu_j)$ by virtue of (6) is equal to the number of elements in the set $X^n\big(O(\mu_j)\big)$, which by (7) does not exceed $e^{n\left(h(\mu_j)+\varepsilon\right)}$. From here, taking into account (12) and (13), we obtain the estimate

$$\sum_i |Z_{n_i}(x_i)|^s \le \sum_{n \ge m}\sum_{j=1}^{l} e^{n\left(h(\mu_j)+\varepsilon\right)} e^{-n\left(h(\mu_j)+2\varepsilon\right)} =$$

$$= \sum_{n \ge m}\sum_{j=1}^{l} e^{-n\varepsilon} = \frac{le^{-m\varepsilon}}{1-e^{-\varepsilon}} \to 0 \text{ as } m \to \infty. \tag{14}$$

It is easy to see that every ball $B(x, r)$ in the space of signals $X^{\mathbb{N}}$, provided $r < 1$, coincides with a cylinder $Z_n(x)$, where n is determined by the conditions $|Z_n(x)| \le r < |Z_{n-1}(x)|$. Therefore every packing of A_N by balls $B(x_i, r_i)$, where $x_i \in A_N$, in fact consists of disjoint cylinders of the form $Z_{n_i}(x_i)$, where

$$|Z_{n_i}(x_i)| \le r_i < \frac{|Z_{n_i}(x_i)|}{\min\limits_j \theta(j)}. \tag{15}$$

It follows from (14), (15) that $C^s(A_N) = 0$. Since the sets A_N cover $\mathrm{B}(W)$, this implies the equality $P^s(\mathrm{B}(W)) = 0$. Hence the packing dimension of the basin $\mathrm{B}(W)$ does not exceed s. In view of arbitrariness of $s > \sup\limits_{\mu \in W} S(\mu, \theta)$ we obtain (4).

Construction of a model set of signals

Let W be a non-empty connected compact subset in $M(X)$. In this section we construct a model set $D_\infty \subset \mathrm{NB}(W)$. We will prove later on that its packing dimension coincides with the dimension of narrow basin $\mathrm{NB}(W)$ specified in theorem 1.

All details of construction of the model set D_∞ are explicitly described in paper [3]. We recall briefly only a general idea of the construction, omitting technical issues, which can be found in [3].

First of all, we need the following lemma.

Lemma 4 [3, lemma 5]. *Let W be a non-empty connected compact subset of a metric space. Then there exists a sequence $x_i \in W$ such that its set of limit points coincides with W and, in addition, $\rho(x_i, x_{i+1}) \to 0$.*

By means of this lemma we choose and fix a sequence of measures $\mu_i \in M(X)$ such that its set of limit points coincides with W and at the same time $\|\mu_i - \mu_{i+1}\| \to 0$. In addition, let all μ_i be strictly positive. This can be ensured by the replacement $\mu_i'(x) = (1 - 2^{-i})\mu_i(x) + \frac{2^{-i}}{k}$ for all $x \in X$ (where $k = |X|$). We preserve the prior notation μ_i for these corrected measures. As a result, the set of limit points of the sequence μ_i will not change (will coincide with W), and the condition $\|\mu_i - \mu_{i+1}\| \to 0$ will remain valid.

Let

$$C = \max\{-\ln\theta(j) \mid j = 1, \ldots, k\}, \tag{16}$$

$$C_i = \max\{-\ln\mu_i(j) \mid j = 1, \ldots, k\}. \tag{17}$$

By positivity of μ_i all the constants C_i are finite.

Choose a sequence of positive numbers ε_i satisfying the condition

$$\varepsilon_i \to 0,\ C_i\varepsilon_i \to 0, \text{ as } i \to \infty. \tag{18}$$

Then construct a sequence of positive integers n_i satisfying the condition

$$n_{i+1} \geq \left(i + \frac{1}{\varepsilon_i}\right) n_i. \tag{19}$$

Using these ε_i and n_i, define the sets

$$A_i = \{x \in X^{n_i} : \|\delta_{x, n_i} - \mu_i\| < \varepsilon_i\} \subset X^{n_i}. \tag{20}$$

Since $\mu_i \in M(X)$ is a probability distribution on X, its Cartesian power $\mu_i^{n_i}$ is a probability distribution on X^{n_i}. The law of large numbers implies that $\mu_i^{n_i}(A_i) \to 1$ as $n_i \to \infty$. Therefore the sequence n_i could be chosen growing so fast that along with (19) the following condition holds true:

$$\frac{n_i^{-1}\left|\ln\mu_i^{n_i}(A_i)\right|}{\min\limits_j \left|\ln\mu_i(j)\right|} \to 0 \text{ as } i \to \infty. \tag{21}$$

Finally, define the model set D_∞ according to the formulae

$$D_i = A_1^{n_2} \times A_2^{n_3} \times \ldots \times A_i^{n_{i+1}}, \tag{22}$$

$$D_\infty = A_1^{n_2} \times A_2^{n_3} \times \ldots. \tag{23}$$

Lemma 5 [3, lemma 6]. *Let W be a non-empty connected compact subset in $M(X)$, and a sequence of strictly positive probability measures $\mu_i \in M(X)$ be such that its set of limit points coincides with W and, at the same time, $\|\mu_i - \mu_{i+1}\| \to 0$. In this setting the model set D_∞ defined in (16)–(23) has the property that for each signal $w \in D_\infty$ the set of limit points of the sequence $\delta_{w, n}$ coincides with W (in other words, $V(w) = W$ and hence $D_\infty \subset \mathrm{NB}(W)$).*

In particular, this lemma implies $\mathrm{NB}(W) \neq \varnothing$.

A lower estimate for the packing dimension of a narrow basin

In this section for the model set D_∞ associated (as described above) with a connected compact subset $W \subset M(X)$ we will prove the estimate

$$\dim_P D_\infty \geq \sup_{\mu \in W} S(\mu, \theta). \tag{24}$$

Inasmuch as $D_\infty \subset \mathrm{NB}(W)$ it will imply that

$$\dim_P \mathrm{NB}(W) \geq \sup_{\mu \in W} S(\mu, \theta). \tag{25}$$

At first notice that for every finite signal $x = (x_1, \ldots, x_n) \in X^n$ and measure $\mu \in M(X)$, we have the relations

$$\mu^n(x) = \mu^{|x|}(x) = \prod_{i=1}^{n} \mu(x_i),$$

$$\ln \mu^{|x|}(x) = \sum_{i=1}^{n} \ln \mu(x_i) = |x| \sum_{j=1}^{k} \delta_{x,n}(j) \ln \mu(j),$$

$$\left| \ln \mu^{|x|}(x) - |x| \sum_{j=1}^{k} \mu(j) \ln \mu(j) \right| \le |x| \cdot \left\| \delta_{x,n} - \mu \right\| \max_{1 \le j \le k} \left| \ln \mu(j) \right|. \tag{26}$$

For every finite signal $x = (x_1, \ldots, x_n)$ we define the cylinder $Z(x)$ consisting of all infinite signals with prefix x:

$$Z(x) = \left\{ w = (w_1, w_2, \ldots) \in X^{\mathbb{N}} \,\middle|\, w_1 = x_1, \ldots, w_n = x_n \right\}.$$

Denote by $|Z(x)|$ its diameter with respect to metric (1). Obviously,

$$|Z(x)| = \prod_{i=1}^{n} \theta(x_i),$$

$$\ln |Z(x)| = \sum_{i=1}^{n} \ln \theta(x_i) = |x| \sum_{j=1}^{k} \delta_{x,n}(j) \ln \theta(j),$$

$$\left| \ln |Z(x)| - |x| \sum_{j=1}^{k} \mu(j) \ln \theta(j) \right| \le |x| \cdot \left\| \delta_{x,n} - \mu \right\| \max_{1 \le j \le k} \left| \ln \theta(j) \right|. \tag{27}$$

Take the sequence of measures $\mu_i \in M(X)$ used in the construction of model set D_∞ and define the following probability measure μ on $X^{\mathbb{N}}$:

$$\mu = \mu_1^{n_1 n_2} \times \mu_2^{n_2 n_3} \times \ldots. \tag{28}$$

For each signal $w \in X^{\mathbb{N}}$ we denote by w' its finite prefixes of arbitrary lengths. Since every ball $B(w, r) \subset X^{\mathbb{N}}$ coincides with a cylinder $Z_n(w)$, where $|Z_n(w)| \le r < |Z_{n-1}(w)|$,

$$D_\mu(x) = \limsup_{r \to 0+0} \frac{\ln \mu(B(x, r))}{\ln r} = \limsup_{|w'| \to \infty} \frac{\ln \mu(Z(w'))}{\ln |Z(w')|}. \tag{29}$$

Further we will prove the estimate

$$\limsup_{|w'| \to \infty} \frac{\ln \mu(Z(w'))}{\ln |Z(w')|} \ge \sup_{\mu \in W} S(\mu, \theta) \text{ for all } w \in D_\infty. \tag{30}$$

If the right-hand side in (30) vanishes then this estimate is trivial. Therefore it is sufficient to consider the case when the right-hand side in (30) is positive.

Fix an arbitrary real number s satisfying the conditions

$$0 < s < \sup_{\mu \in W} S(\mu, \theta).$$

The function $S(\mu, \theta)$ attains its maximum on the compact set W at a certain point $\mu^* \in W$. By construction μ^* is a limit point for the sequence of measures μ_i. Hence there exists an infinite subset $I \subset \mathbb{N}$ such that for all $i \in I$,

$$S(\mu_i, \theta) = \frac{\sum_{j=1}^{k} \mu_i(j) \ln \mu_i(j)}{\sum_{j=1}^{k} \mu_i(j) \ln \theta(j)} > s$$

and, consequently,

$$\sum_{j=1}^{k} \mu_i(j) \ln \mu_i(j) < s \sum_{j=1}^{k} \mu_i(j) \ln \theta(j), \ i \in I. \tag{31}$$

Take a number $i \in I$. Consider an arbitrary model signal $w \in D_\infty$. By construction it has prefix of the form $w' = xy$, where $x \in D_{i-1}$ and $y \in A_i^{n_{i+1}}$.
Notice that in view of (19), (20), (22)

$$\frac{|x|}{|y|} = \frac{n_1 n_2 + \ldots + n_{i-1} n_i}{n_i n_{i+1}} \le \frac{(i-1) n_{i-1} n_i}{n_i n_{i+1}} \le \frac{n_i}{n_{i+1}} \le \varepsilon_i. \tag{32}$$

Let us estimate the value of $\ln \mu\left(Z(w')\right)$. By definition (28) of μ,

$$\mu\left(Z(w')\right) \le \mu_i^{|y|}(y). \tag{33}$$

It follows from (26) that

$$\ln \mu_i^{|y|}(y) \le |y| \sum_{j=1}^{k} \mu_i(j) \ln \mu_i(j) + |y| \cdot \left\| \delta_{y,|y|} - \mu_i \right\| \max_{1 \le j \le k} \left| \ln \mu_i(j) \right|. \tag{34}$$

Combining (33), (34), the inequality $\left\| \delta_{y,|y|} - \mu_i \right\| < \varepsilon_i$ (which follows from definition (20) of the set A_i), and equality (17), having the form $\max\limits_j \left| \ln \mu_i(j) \right| = C_i$, we obtain the estimate

$$\ln \mu\left(Z(w')\right) \le \ln \mu_i^{|y|}(y) \le |y| \sum_{j=1}^{k} \mu_i(j) \ln \mu_i(j) + |y| \varepsilon_i C_i. \tag{35}$$

Then, substitution of (31) in (35) gives

$$\ln \mu\left(Z(w')\right) \le s|y| \sum_{j=1}^{k} \mu_i(j) \ln \theta(j) + |y| \varepsilon_i C_i. \tag{36}$$

The product $C_i \varepsilon_i$ in view of (18) tends to zero. Therefore the second summand in the right-hand side of (36) is infinitely small with respect to the first one. So (36) can be written in the form

$$\ln \mu\left(Z(w')\right) \le s \left(|y| \sum_{j=1}^{k} \mu_i(j) \ln \theta(j) \right) \left(1 + \alpha_i(w)\right), \ i \in I, \tag{37}$$

where $\alpha_i(w) \to 0$ as $i \to \infty$.

In the same way we may estimate $\ln \left| Z(w') \right|$ from below by means of (16), (27):

$$\ln \left| Z(w') \right| = \ln \left| Z(x) \right| + \ln \left| Z(y) \right| \ge -|x| C + |y| \sum_{j=1}^{k} \mu_i(j) \ln \theta(j) - |y| \varepsilon_i C. \tag{38}$$

Recall that $\varepsilon_i \to 0$, and from (32) it follows $|x| \le \varepsilon_i |y|$. Therefore the first and third summands in the right-hand side of (38) are infinitely small with respect to the second one, and so (38) can be written in the form

$$\ln \left| Z(w') \right| \ge \left(|y| \sum_{j=1}^{k} \mu_i(j) \ln \theta(j) \right) \left(1 + \beta_i(w)\right), \tag{39}$$

where $\beta_i(w) \to 0$ as $i \to \infty$.

Dividing (37) by (39) and taking into account that the left and right hand sides in these inequalities are negative, we obtain

$$\frac{\ln \mu\left(Z(w')\right)}{\ln \left| Z(w') \right|} \ge s \frac{1 + \alpha_i(w)}{1 + \beta_i(w)}, \ i \in I.$$

It follows from here that for each model signal $w \in D_\infty$ and its prefixes w',

$$\limsup_{|w'|\to\infty}\frac{\ln\mu\left(Z\left(w'\right)\right)}{\ln\left|Z\left(w'\right)\right|}\geq s. \tag{40}$$

In view of arbitrariness of the number $s<\sup\limits_{\mu\in W}S\left(\mu,\theta\right)$ the last inequality implies (30).

If $\mu\left(D_\infty\right)>0$ then (24) follows from (29), (30), and theorem 3. But in fact, the equality $\mu\left(D_\infty\right)=0$ is most likely to take place. In this case it is enough to replace the measure μ in (30) by a probability measure ν on D_∞ such that for any signal $w\in D_\infty$,

$$\lim_{|w'|\to\infty}\frac{\ln\nu\left(Z\left(w'\right)\right)}{\ln\mu\left(Z\left(w'\right)\right)}=1. \tag{41}$$

To this end we define measures ν_i on the alphabet X by the formula

$$\nu_i=\frac{\mu_i}{\left(\mu_i^{n_i}\left(A_i\right)\right)^{1/n_i}} \tag{42}$$

and a measure ν on the model set $D_\infty=A_1^{n_2}\times A_2^{n_3}\times\ldots$ (of the same type as in (28)):

$$\nu=\nu_1^{n_1n_2}\times\nu_2^{n_2n_3}\times\ldots.$$

By construction $\nu_i^{n_i}\left(A_i\right)=1$. Therefore $\nu\left(D_\infty\right)=1$. Formally ν is not defined outside the model set D_∞ but it may be extended by zero if one wishes.

It follows from (21), (42) that when $i\to\infty$,

$$\frac{\left|\ln\nu_i\left(j\right)\right|}{\left|\ln\mu_i\left(j\right)\right|}=\left|1-\frac{n_i^{-1}\ln\mu_i^{n_i}\left(A_i\right)}{\ln\mu_i\left(j\right)}\right|\to 1\text{ uniformly on }\ j\in X. \tag{43}$$

Evidently, $\left|Z\left(w'\right)\right|\to 0$ as $\left|w'\right|\to\infty$. From here and (40) it follows that $\mu\left(Z\left(w'\right)\right)\to 0$ and hence $\left|Z\left(w'\right)\right|\to 0$ as $\left|w'\right|\to\infty$. The last convergence along with (43) implies equality (41). Thus estimates (24), (25) are completely proved.

The union of estimates (4) and (25) looks as follows:

$$\sup_{\mu\in W}S\left(\mu,\theta\right)\leq\dim_P\mathrm{NB}\left(W\right)\leq\dim_P\mathrm{B}\left(W\right)\leq\sup_{\mu\in W}S\left(\mu,\theta\right), \tag{44}$$

where the left inequality is proved for the connected compact subsets $W\subset M\left(X\right)$ and the right one for all non-empty subsets $W\subset M\left(X\right)$. This immediately implies theorem 1.

To prove equality (2) from theorem 2, it is sufficient to notice that in view of (44) for any measure $\mu^*\in W$ we have

$$S\left(\mu^*,\theta\right)\leq\dim_P B\left(\mu^*\right)\leq\dim_P\mathrm{B}\left(W\right)\leq\sup_{\mu\in W}S\left(\mu,\theta\right),$$

and take supremum over $\mu^*\in W$.

Equality (3) from theorem 2 follows from (44) and the inclusions

$$\mathrm{NB}\left(M\left(X\right)\right)\subset\mathrm{WB}\left(W\right)\subset\mathrm{B}\left(M\left(X\right)\right)=X^{\mathbb{N}}.$$

Библиографические ссылки

1. Billingsley P. Hausdorff dimension in probability theory. *Illinois Journal of Mathematics.* 1960;4:187–209.
2. Billingsley P. Hausdorff dimension in probability theory II. *Illinois Journal of Mathematics.* 1961;5:291–298.
3. Бахтин ВИ, Садок БМ. Хаусдорфовы размерности узких бассейнов в пространстве последовательностей. *Труды Института математики.* 2019;27(1–2):3–12.
4. Бахтин ВИ, Садок Б. Упаковочные размерности бассейнов в пространстве последовательностей. *Доклады Национальной академии наук Беларуси.* 2020;64(3):263–267. DOI: 10.29235/1561-8323-2020-64-3-263-267.
5. Bakhtin V. The McMillan theorem for colored branching processes and dimensions of random fractals. *Entropy.* 2014;16(12): 6624–6653. DOI: 10.3390/e16126624.
6. Falconer F. *Techniques in fractal geometry.* Chichester: John Wiley & Sons; 1997. [260 p.].
7. Shiryaev AN. *Probability-1.* 3rd edition. [S. l.]: Springer; 2016. 503 p.

References

1. Billingsley P. Hausdorff dimension in probability theory. *Illinois Journal of Mathematics.* 1960;4:187–209.
2. Billingsley P. Hausdorff dimension in probability theory II. *Illinois Journal of Mathematics.* 1961;5:291–298.
3. Bakhtin VI, Sadok BM. Hausdorff dimensions of narrow basins in the space of sequences. *Proceedings of the Institute of Mathematics.* 2019;27(1–2):3–12. Russian.
4. Bakhtin VI, Sadok B. [Packing dimensions of basins in the space of sequences]. *Doklady Natsional'noi akademii nauk Belarusi.* 2020;64(3):263–267. Russian. DOI: 10.29235/1561-8323-2020-64-3-263-267.
5. Bakhtin V. The McMillan theorem for colored branching processes and dimensions of random fractals. *Entropy.* 2014;16(12): 6624–6653. DOI: 10.3390/e16126624.
6. Falconer F. *Techniques in fractal geometry.* Chichester: John Wiley & Sons; 1997. [260 p.].
7. Shiryaev AN. *Probability-1.* 3rd edition. [S. l.]: Springer; 2016. 503 p.

Received 20.04.2021 / revised 07.07.2021 / accepted 07.07.2021.

УДК 517.968.7

РЕШЕНИЕ ОДНОГО ГИПЕРСИНГУЛЯРНОГО ИНТЕГРО-ДИФФЕРЕНЦИАЛЬНОГО УРАВНЕНИЯ, ЗАДАННОГО С ПОМОЩЬЮ ОПРЕДЕЛИТЕЛЕЙ

А. П. ШИЛИН[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Приводится точное аналитическое решение гиперсингулярного интегро-дифференциального уравнения произвольного порядка на замкнутой кривой, расположенной в комплексной плоскости. Характерная особенность уравнения состоит в том, что оно записано с помощью определителей. С точки зрения традиционной классификации уравнений его следует отнести к линейным уравнениям с переменными коэффициентами специального вида. Применяется метод аналитического продолжения. Уравнение сводится к краевой задаче линейного сопряжения для аналитических функций с некоторыми дополнительными условиями. В случае разрешимости этой задачи требуется решить еще два линейных дифференциальных уравнения в классе аналитических функций. Указываются в явном виде условия разрешимости, при выполнении которых решение также может быть записано явно. Рассматривается пример.

Ключевые слова: интегро-дифференциальные уравнения; гиперсингулярные интегралы; обобщенные формулы Сохоцкого; дифференциальные уравнения; краевая задача Римана.

SOLUTION OF ONE HYPERSINGULAR INTEGRO-DIFFERENTIAL EQUATION DEFINED BY DETERMINANTS

A. P. SHILIN[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

The paper provides an exact analytical solution to a hypersingular integro-differential equation of arbitrary order. The equation is defined on a closed curve in the complex plane. A characteristic feature of the equation is that if is written using determinants. From the view of the traditional classification of the equations, it should be classified as linear equations with variable coefficients of a special form. The method of analytical continuation id applied. The equation is reduced to a boundary value problem of linear conjugation for analytic functions with some additional conditions. If this problem is solvable, if is required to solve two more linear differential equations in the class of analytic functions. The conditions of solvability are indicated explicitly. When these conditions are met, the solution can also be written explicitly. An example is given.

Keywords: integro-differential equations; hypersingular integrals; generalised Sokhotsky formulas; differential equations; Riemann boundary problem.

Образец цитирования:
Шилин АП. Решение одного гиперсингулярного интегро-дифференциального уравнения, заданного с помощью определителей. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:17–28.
https://doi.org/10.33581/2520-6508-2021-2-17-28

For citation:
Shilin AP. Solution of one hypersingular integro-differential equation defined by determinants. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:17–28. Russian.
https://doi.org/10.33581/2520-6508-2021-2-17-28

Автор:
Андрей Петрович Шилин – кандидат физико-математических наук, доцент; доцент кафедры высшей математики и математической физики физического факультета.

Author:
Andrei P. Shilin, PhD (physics and mathematics), docent; associate professor at the department of higher mathematics and mathematical physics, faculty of physics.
a.p.shilin@gmail.com

Введение

В гиперсингулярных интегральных уравнениях (ГИУ) интегралы понимаются в смысле конечной части по Адамару. Основными методами решения таких уравнений являются численные методы, представление о которых можно получить, например, в статье [1]. Аналитические методы решения разработаны мало. Достаточно подробный обзор современного состояния методов решения ГИУ дан в работе [2], где, в частности, сказано: «В настоящее время ГИУ находят широкое применение при моделировании задач аэродинамики, электродинамики, микроэлектроники, геофизики, атомной и ядерной физики и ряда других областей естествознания и техники» [2, с. 245].

Точное аналитическое решение гиперсингулярного интегро-дифференциального уравнения впервые, по-видимому, приведено в статье [3] – это решение линейного уравнения на замкнутой кривой в комплексной плоскости в случае постоянных коэффициентов. В настоящей работе, как и в публикациях [4–6], аналогичное уравнение рассматривается для таких случаев переменных коэффициентов, когда возможность точного аналитического решения сохраняется.

Постановка задачи

Обозначим через L простую гладкую замкнутую кривую на расширенной комплексной плоскости. Пусть $D_{\pm}$ – области, для которых кривая L является границей, $0 \in D_{+}$, $\infty \in D_{-}$. Ориентируем кривую L так, чтобы при движении по ней в положительном направлении область D_{+} оставалась слева.

Пусть $m, n \in \mathbb{N}$. Зададим H-непрерывные (т. е. удовлетворяющие условию Гёльдера) функции $a(t) \neq 0$, $b(t) \neq 0$, $f(t)$, m раз H-непрерывно дифференцируемые функции $p_j(t)$, $j = \overline{1, m}$, и n раз H-непрерывно дифференцируемые функции $q_j(t)$, $j = \overline{1, n}$, $t \in L$. На кривой L требуется найти $\max(m, n)$ раз H-непрерывно дифференцируемую функцию $\varphi(t)$, удовлетворяющую уравнению

$$a(t)\begin{vmatrix} p_1(t) & p_2(t) & \dots & p_m(t) & \varphi(t) \\ p_1'(t) & p_2'(t) & \dots & p_m'(t) & \varphi'(t) \\ \dots & \dots & \dots & \dots & \dots \\ p_1^{(m)}(t) & p_2^{(m)}(t) & \dots & p_m^{(m)}(t) & \varphi^{(m)}(t) \end{vmatrix} + b(t)\begin{vmatrix} q_1(t) & q_2(t) & \dots & q_n(t) & \varphi(t) \\ q_1'(t) & q_2'(t) & \dots & q_n'(t) & \varphi'(t) \\ \dots & \dots & \dots & \dots & \dots \\ q_1^{(n)}(t) & q_2^{(n)}(t) & \dots & q_n^{(n)}(t) & \varphi^{(n)}(t) \end{vmatrix} +$$

$$+ \frac{a(t)}{\pi i}\begin{vmatrix} p_1(t) & p_2(t) & \dots & p_m(t) & 0!\int\limits_L \dfrac{\varphi(\tau)d\tau}{\tau - t} \\ p_1'(t) & p_2'(t) & \dots & p_m'(t) & 1!\int\limits_L \dfrac{\varphi(\tau)d\tau}{(\tau - t)^2} \\ \dots & \dots & \dots & \dots & \dots \\ p_1^{(m)}(t) & p_2^{(m)}(t) & \dots & p_m^{(m)}(t) & m!\int\limits_L \dfrac{\varphi(\tau)d\tau}{(\tau - t)^{m+1}} \end{vmatrix} -$$

$$- \frac{b(t)}{\pi i}\begin{vmatrix} q_1(t) & q_2(t) & \dots & q_n(t) & 0!\int\limits_L \dfrac{\varphi(\tau)d\tau}{\tau - t} \\ q_1'(t) & q_2'(t) & \dots & q_n'(t) & 1!\int\limits_L \dfrac{\varphi(\tau)d\tau}{(\tau - t)^2} \\ \dots & \dots & \dots & \dots & \dots \\ q_1^{(n)}(t) & q_2^{(n)}(t) & \dots & q_n^{(n)}(t) & n!\int\limits_L \dfrac{\varphi(\tau)d\tau}{(\tau - t)^{n+1}} \end{vmatrix} = f(t),\ t \in L, \qquad (1)$$

с интегралами, понимаемыми в смысле конечной части по Адамару.

Будем обозначать через W вронскиан функций, указывая в скобках сами функции и их аргумент. Для $t \in L$ введем обозначения

$$W_{+}(t) = W(p_1, p_2, \dots, p_m; t),\ W_j^{+}(t) = W(p_1, \dots, p_{j-1}, p_{j+1}, \dots, p_m; t),\ j = \overline{1, m},$$

$$W_{-}(t) = W(q_1, q_2, \dots, q_n; t),\ W_j^{-}(t) = W(q_1, \dots, q_{j-1}, q_{j+1}, \dots, q_n; t),\ j = \overline{1, n}.$$

Пусть точка z_0 является для некоторой функции точкой аналитичности либо полюсом. Напомним, что порядком этой функции в точке z_0 называется число $\nu \in \mathbb{Z}$ из разложения функции в ряд $\sum\limits_{k=\nu}^{\infty} c_k\left(z-z_0\right)^k$, $c_\nu \neq 0$, в окрестности этой точки.

Возможность решить уравнение (1) анонсирована в статье [7]. В настоящей работе приведем подробно соответствующий результат, допуская к тому же наличие у исходных функций особых точек, значительно усложняющих решение. Как и в статье [7], в дальнейшем будем считать, что все функции $p_j(t)$ (а следовательно, и $W_+(t)$, $W_j^+(t)$) аналитически продолжимы в D_+, причем $W_+(z) \neq 0$, $z \in D_+ \cup L$. Но сделаем исключение для одной (особой) точки $z_* \in D_+$: пусть в этой точке порядки функций $p_j(z)$ равны $k_j \in \mathbb{Z}$ (и тогда необязательно $W_+(z_*) \neq 0$). Как и в работе [7], будем предполагать также, что все функции $q_j(t)$ (а значит, и $W_-(t)$, $W_j^-(t)$) аналитически продолжимы в D_-, причем $W_-(z) \neq 0$, $z \in (D_- \cup L)\setminus\{\infty\}$. Роль особой точки в D_- будет играть точка $z = \infty$, поскольку всегда $W_-(\infty) = 0$. В статье [7] изучен случай, когда $\Delta \neq 0$, где

$$\Delta = \begin{vmatrix} k_{10} & k_{20} & \dots & k_{n0} \\ k_{11} & k_{21} & \dots & k_{n1} \\ \dots & \dots & \dots & \dots \\ k_{1,\,n-1} & k_{2,\,n-1} & \dots & k_{n,\,n-1} \end{vmatrix},$$

а элементы определителя Δ берутся из разложений в ряды Тейлора $q_j(z) = \sum\limits_{s=0}^{\infty} \frac{k_{js}}{z^s}$ функций $q_j(z)$ в окрестности точки $z = \infty$. Теперь предположим, что функции $q_j(z)$ имеют в точке $z = \infty$ нули некоторых порядков $l_j \in \mathbb{N}$, $j = \overline{1,\,n}$ (и тогда всегда будет $\Delta = 0$).

Основной результат

Лемма 1. *Если порядки k_j функций $p_j(z)$, $j = \overline{1,\,m}$, в точке z_* попарно различны, то порядок $W_+(z)$ в этой точке равен* $\sum\limits_{j=1}^{m} k_j - \frac{(m-1)m}{2}$.

Доказательство. Разложим в точке z_* элементы определителя $W_+(z)$ в ряды Лорана (в частности, это могут быть ряды Тейлора):

$$\begin{vmatrix} a_1(z-z_*)^{k_1} + \dots & \dots & a_m(z-z_*)^{k_m} + \dots \\ a_1k_1(z-z_*)^{k_1-1} + \dots & \dots & a_mk_m(z-z_*)^{k_m-1} + \dots \\ a_1k_1(k_1-1)(z-z_*)^{k_1-2} + \dots & \dots & a_mk_m(k_m-1)(z-z_*)^{k_m-2} + \dots \\ a_1k_1(k_1-1)(k_1-2)(z-z_*)^{k_1-3} + \dots & \dots & a_mk_m(k_m-1)(k_m-2)(z-z_*)^{k_m-3} + \dots \\ \dots & \dots & \dots \\ a_1k_1(k_1-1)(k_1-2)\dots(k_1-m+2)(z-z_*)^{k_1-m+1} + \dots & \dots & a_mk_m(k_m-1)(k_m-2)\dots(k_m-m+2)(z-z_*)^{k_m-m+1} + \dots \end{vmatrix}.$$

В каждом элементе определителя указаны лишь начальные слагаемые соответствующего ряда, остальные слагаемые обозначены многоточиями, a_1, …, a_m – ненулевые постоянные. Подчеркнем, что при дифференцировании членов рядов мы пользуемся формулой вида $\left((z-z_*)^s\right)^{(k)} = s(s-1)\dots(s-k+1)(z-z_*)^{s-k}$ в том числе и при $0 \le s < k$, когда $\left((z-z_*)^s\right)^{(k)} = 0$. Вынесем за знак определителя множитель $a_j(z-z_*)^{k_j-m+1}$ из j-го столбца, затем множитель $(z-z_*)^{m-j}$ из j-й строки, $j = \overline{1,\,m}$. В результате перед определителем будет множитель $A(z-z_*)^M$, где $A = a_1a_2\cdots a_m \neq 0$, $M = \sum\limits_{j=1}^{m}(k_j-m+1) + \sum\limits_{j=1}^{m}(m-j) = \sum\limits_{j=1}^{m} k_j - \frac{(m-1)m}{2}$, а определитель после раскрытия скобок и приведения подобных членов примет вид

$$\begin{vmatrix} 1+\ldots & \ldots & 1+\ldots \\ k_1+\ldots & \ldots & k_m+\ldots \\ k_1^2-k_1+\ldots & \ldots & k_m^2-k_m+\ldots \\ k_1^3-3k_1^2+2k_1+\ldots & \ldots & k_m^3-3k_m^2+2k_m+\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ k_1^{m-1}+\ldots+(-1)^m(m-2)!k_1+\ldots & \ldots & k_m^{m-1}+\ldots+(-1)^m(m-2)!k_m+\ldots \end{vmatrix},$$

где все многоточия в элементах определителя обозначают бесконечно малые функции при $z \to z_*$.

Теперь полученный определитель представим в виде надлежащей суммы таких определителей, элементами которых будут отдельные слагаемые строк. Ненулевым будет лишь определитель из всех первых слагаемых строк, поскольку он совпадает с определителем V Вандермонда попарно различных чисел $k_1, k_2, \ldots, k_m$. Остальные определители или обратятся в ноль из-за пропорциональности каких-либо строк, или будут бесконечно малыми при $z \to z_*$ из-за наличия таких бесконечно малых хотя бы в одной строке.

Итак,

$$W_+(z) = AV(z-z_*)^M + o\left((z-z_*)^M\right), \ z \to z_*.$$

Лемма доказана.

Отметим, что при выполнении условий леммы порядки $W_j^+(z)$ в точке z_*, очевидно, будут равны $\sum\limits_{\substack{s=1 \\ s\neq j}}^{m} k_s - \dfrac{(m-2)(m-1)}{2}$, $j = \overline{1, m}$.

Проанализируем для определителя $W_+(z)$ случай произвольных целых порядков k_j, $j = \overline{1, m}$. Обозначим $\tilde{k}_1 = k_{j_1}$ наименьший из этих порядков. Если наряду с функцией $p_{j_1}(z)$ некоторые из функций $p_j(z)$ также имеют этот порядок, то прибавим к столбцам определителя $W_+(z)$, содержащим такие функции, столбец с функцией $p_{j_1}(z)$, умноженный на надлежащие константы, так, чтобы в результате столбец с номером j_1 остался единственным, в котором порядок первой функции равен $\tilde{k}_1$. Обозначим $\tilde{p}_j(z)$ функции в первой строке нового определителя, $j = \overline{1, m}$. В частности, будет $\tilde{p}_{j_1}(z) \equiv p_{j_1}(z)$. Если функция $p_{j_1}(z)$ окажется единственной с порядком $\tilde{k}_1$ в точке z_*, то просто считаем $\tilde{p}_j(z) \equiv p_j(z)$, $j = \overline{1, m}$.

Теперь в новом определителе обозначим $\tilde{k}_2$ наименьший порядок в точке z_* всех функций $\tilde{p}_j(z)$, кроме функции $\tilde{p}_{j_1}(z)$, и пусть функция $\tilde{p}_{j_2}(z)$ имеет этот порядок. Если некоторые из функций $\tilde{p}_j(z)$ также имеют этот порядок в точке z_*, то прибавим к столбцам с такими функциями столбец с функцией $\tilde{p}_{j_2}(z)$, умноженный на надлежащие константы, так, чтобы в результате столбец с номером j_2 остался единственным, в котором порядок первой функции равен $\tilde{k}_2$. Обозначим $\tilde{\tilde{p}}_j(z)$ функции в первой строке нового определителя. Если функция $\tilde{p}_{j_2}(z)$ будет единственной с порядком $\tilde{k}_2$ в точке z_*, то просто считаем $\tilde{\tilde{p}}_j(z) \equiv \tilde{p}_j(z)$ для всех $j = \overline{1, m}$. Обозначим $\tilde{k}_3$ наименьший порядок в точке z_* всех функций $\tilde{\tilde{p}}_j(z)$, кроме функций $\tilde{\tilde{p}}_{j_1}(z)$ и $\tilde{\tilde{p}}_{j_2}(z)$, и продолжим действия со столбцами, аналогичные предыдущим. В результате мы добьемся того, что порядки в точке z_* всех функций в первой строке определителя станут попарно различными. Поскольку значение определителя в результате указанных действий со столбцами не изменится, то можно, не теряя общности, считать с самого начала все порядки k_j попарно различными, что мы и будем предполагать в дальнейших рассуждениях.

Лемма 2. *Если порядки l_j нуля на бесконечности функций $q_j(z)$, $j = \overline{1, n}$, попарно различны, то порядок нуля на бесконечности $W_-(z)$ равен* $\sum\limits_{j=1}^{n} l_j + \dfrac{(n-1)n}{2}$.

Подробное доказательство не приводится, поскольку оно аналогично доказательству леммы 1. Отметим только, что, прежде чем проводить похожие рассуждения, элементы первой строки определителя $W_-(z)$ следует представить в виде $b_j z^{-l_j} + \ldots$, $b_j \in \mathbb{C}$, $b_j \neq 0$, $j = \overline{1, n}$. Получим, что $W_-(z)$ относительно z будет иметь степень на бесконечности $\sum\limits_{j=1}^{n}(-l_j) - \dfrac{(n-1)n}{2}$, т. е. точка $z = \infty$ будет нулем порядка

$\sum_{j=1}^{n} l_j + \frac{(n-1)n}{2}$. Укажем также, что для определителей $W_j^-(z)$ нуль на бесконечности будет иметь порядки $\sum_{\substack{s=1 \\ s\neq j}}^{n} l_s + \frac{(n-2)(n-1)}{2}$, $j=\overline{1,\, n}$. Так же как и в случае определителя $W_+(z)$, в дальнейшем можно, не теряя общности, всегда считать для определителя $W_-(z)$ все порядки l_j попарно различными.

Введем новые неизвестные функции:

$$\Phi_\pm(z)=\frac{1}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{\tau - z},\ z\in D_\pm,$$

$$F_+(z)=W\left(p_1,\, p_2,\, \ldots,\, p_m,\, \Phi_+;\, z\right),\ z\in D_+, \tag{2}$$

$$F_-(z)=W\left(q_1,\, q_2,\, \ldots,\, q_n,\, \Phi_-;\, z\right),\ z\in D_-. \tag{3}$$

В определителе $W\left(p_1,\, p_2,\, \ldots,\, p_m,\, \Phi_+;\, z\right)$ без потери общности рассуждений мы можем считать порядки функций $p_j(z)$ в точке z_* попарно различными и равными k_j, $j=\overline{1,\, m}$. В противном случае к этому можно прийти теми же действиями с первыми m столбцами, что и в определителе $W_+(z)$: данные действия не изменят определитель точно так же, как и определитель $W_+(z)$. Аналогичное замечание справедливо для определителя в уравнении (3).

Функции $\Phi_\pm(z)$ аналитичны в соответствующих областях $D_\pm$, $\Phi_-(\infty)=0$. Функция $F_+(z)$ аналитична в D_+, кроме, возможно, точки $z=z_*$, где у нее может быть полюс. Функция $F_-(z)$ аналитична в D_-, $F_-(\infty)=0$.

Если порядок функции $\Phi_+(z)$ в точке z_* не совпадает ни с одним из порядков k_j функций $p_j(z)$, то порядок α в этой точке у функции $F_+(z)$ можно вычислить, применяя лемму 1 к функциям $p_1(z),\ p_2(z),\ \ldots,\ p_m(z),\ \Phi_+(z)$. Для определенности в дальнейшем будем полагать $k_j\neq 0$, $j=\overline{1,\, m}$. При этом минимально возможный порядок $\alpha=\sum_{j=1}^{m} k_j - \frac{m(m+1)}{2}$ будет в случае, когда у функции $\Phi_+(z)$ нулевой порядок. Если порядок функции $\Phi_+(z)$ в точке z_* совпадает с одним из порядков k_j, то для расчета порядка в этой точке у функции $F_+(z)$ используем описанную ранее процедуру действий со столбцами $F_+(z)$. При этом допускается изменять лишь последний столбец, а другие оставлять неизменными, из-за чего порядок первого элемента в последнем столбце будет увеличиваться. Функция $F_+(z)$ при этом не изменится, а ее порядок в точке z_* окажется больше α. Таким образом, $F_+(z)=O\left(\left(z-z_*\right)^{\alpha}\right)$, $z\to z_*$. Аналогично, полагая в дальнейшем $l_j\neq 1$, $j=\overline{1,\, n}$, с помощью леммы 2 получим $F_-(z)=O\left(\left(\frac{1}{z}\right)^{\beta}\right)$, $z\to\infty$, где $\beta=\sum_{j=1}^{n} l_j + \frac{n(n+1)}{2}+1$ (единица, в сравнении с числом α, добавится из-за минимально возможного 1-го порядка $F_-(z)$ на бесконечности). Используя терминологию [8], можно сказать, что кусочно-аналитические функции $F_+(z)$, $F_-(z)$ должны быть кратны дивизору $z_*^{\alpha}\infty^{\beta}$.

Вернемся к уравнению (1). Его можно сначала записать в виде

$$a(t)\begin{vmatrix} p_1(t) & p_2(t) & \ldots & p_m(t) & \frac{1}{2}\varphi(t)+\frac{0!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{\tau-t} \\ p_1'(t) & p_2'(t) & \ldots & p_m'(t) & \frac{1}{2}\varphi'(t)+\frac{1!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ p_1^{(m)}(t) & p_2^{(m)}(t) & \ldots & p_m^{(m)}(t) & \frac{1}{2}\varphi^{(m)}(t)+\frac{m!}{2\pi i}\int_L \frac{\varphi(\tau)d\tau}{(\tau-t)^{m+1}} \end{vmatrix} =$$

$$= b(t)\begin{vmatrix} q_1(t) & q_2(t) & \dots & q_n(t) & -\frac{1}{2}\varphi(t) + \frac{0!}{2\pi i}\int\limits_L \frac{\varphi(\tau)d\tau}{\tau - t} \\ q_1'(t) & q_2'(t) & \dots & q_n'(t) & -\frac{1}{2}\varphi'(t) + \frac{1!}{2\pi i}\int\limits_L \frac{\varphi(\tau)d\tau}{(\tau - t)^2} \\ \dots & \dots & \dots & \dots & \dots \\ q_1^{(n)}(t) & q_2^{(n)}(t) & \dots & q_n^{(n)}(t) & -\frac{1}{2}\varphi^{(n)}(t) + \frac{n!}{2\pi i}\int\limits_L \frac{\varphi(\tau)d\tau}{(\tau - t)^{n+1}} \end{vmatrix} + \frac{f(t)}{2},\ t \in L,$$

а потом, используя обобщенные формулы Сохоцкого [9], в виде краевой задачи Римана

$$F_+(t) = \frac{b(t)}{a(t)}F_-(t) + \frac{f(t)}{2a(t)},\ t \in L. \tag{4}$$

Если решение этой задачи будет найдено, то соотношения (2), (3) следует затем расценивать как линейные дифференциальные уравнения для нахождения функций $\Phi_\pm(z)$. В случае если и функции $\Phi_\pm(z)$ будут найдены, решение уравнения (1) можно найти по формуле

$$\varphi(t) = \Phi_+(t) - \Phi_-(t),\ t \in L. \tag{5}$$

Таким образом, справедлива следующая теорема.

Теорема. *Решение уравнения (1) сводится к последовательному решению краевой задачи Римана (4) в классе функций, кратных дивизору $z_*^{\alpha}\infty^{\beta}$, и дифференциальных уравнений (2), (3) в классе аналитических функций (с условием $\Phi_-(\infty) = 0$).*

Дальнейшее исследование

Решение задачи Римана. Теория задачи Римана [10] позволяет решить возникшую задачу (4). Введем следующие обозначения: $\varkappa = \operatorname{Ind}_L \frac{b(t)}{a(t)}$; $X_\pm(z)$ – канонические функции задачи Римана, факторизующие ее коэффициент $\frac{b(t)}{a(t)}$; $\Psi_\pm(z) = \frac{1}{4\pi i}\int\limits_L \frac{f(\tau)d\tau}{a(\tau)X_+(\tau)(\tau - z)}$, $z \in D_\pm$. Решение задачи (4) имеет вид

$$F_\pm(z) = X_\pm(z)\left(P_{\varkappa-\beta}(z) + \frac{\Psi_\pm(z)(z - z_*)^{|\alpha|} + Q_{|\alpha|-1}(z)}{(z - z_*)^{|\alpha|}}\right),\ z \in D_\pm.$$

При $\varkappa - \beta < 0$ $P_{\varkappa-\beta}(z) \equiv 0$, а при $\varkappa - \beta \geq 0$ $P_{\varkappa-\beta}(z)$ – многочлен вида $\sum\limits_{k=0}^{\varkappa-\beta} c_k(z - z_*)^k$, коэффициенты которого зависят от величины α:

- в случае $\alpha \leq 0$ являются произвольными;
- в случае $\alpha > \varkappa - \beta$ вычисляются по формулам

$$c_k = \frac{(-1)^{k+1}}{4\pi i}\int\limits_L \frac{f(\tau)d\tau}{a(\tau)(\tau - z_*)^{k+1}} \tag{6}$$

для всех $k = \overline{0,\ \varkappa - \beta}$;

- в случае $1 \leq \alpha \leq \varkappa - \beta$ вычисляются по формулам (6) для $k = \overline{0,\ \alpha - 1}$, а для $k = \overline{\alpha,\ \varkappa - \beta}$ остаются произвольными.

Если $\alpha \geq 0$, то $Q_{|\alpha|-1}(z) \equiv 0$, если же $\alpha \leq -1$, то $Q_{|\alpha|-1}(z)$ – многочлен вида $\sum\limits_{k=0}^{|\alpha|-1} d_k(z - z_*)^k$, коэффициенты которого зависят от величины $\varkappa - \beta$:

- в случае $\varkappa - \beta \geq -1$ являются произвольными;
- в случае $\varkappa - \beta \leq -|\alpha| - 1$ вычисляются по формулам

$$d_{|\alpha|-k}=\frac{1}{4\pi i}\int_L \frac{f(\tau)(\tau-z_*)^{k-1}d\tau}{a(\tau)} \tag{7}$$

для всех $k=\overline{1,|\alpha|}$;

• в случае $-|\alpha|\le \varkappa-\beta\le -2$ вычисляются по формулам (7) для $k=\overline{1,\beta-\varkappa-1}$, а для $k=\overline{\beta-\varkappa,|\alpha|}$ остаются произвольными.

При этом для разрешимости задачи (4) необходимым и достаточным является:

1) при $1\le\varkappa-\beta+1<\alpha$ выполнение условий

$$\int_L \frac{f(\tau)d\tau}{a(\tau)X_+(\tau)(\tau-z_*)^{k+1}}=0, \tag{8}$$

в которых $k=\overline{\varkappa-\beta+1,\alpha-1}$;

2) при $\varkappa-\beta<-|\alpha|-1$, $\alpha\le -1$ выполнение условий

$$\int_L \frac{f(\tau)(\tau-z_*)^{k-1}d\tau}{a(\tau)}=0, \tag{9}$$

в которых $k=\overline{|\alpha|+1,\beta-\varkappa-1}$;

3) при $\varkappa-\beta<-1$, $\alpha>0$ выполнение совокупности условий (8), в которых $k=\overline{0,\alpha-1}$, и условий (9), в которых $k=\overline{1,\beta-\varkappa-1}$.

В остальных случаях, т. е. при $\varkappa-\beta-\alpha\ge -1$, задача разрешима безусловно.

Решение дифференциальных уравнений. Предполагая, что задача Римана (4) разрешима, а ее решение найдено, будем исследовать в области D_+ уравнение (2). Решение этого уравнения, полученное методом вариации произвольных постоянных, имеет вид

$$\Phi_+(z)=\sum_{j=1}^{m}\left(C_j^+p_j(z)+p_j(z)\tilde{C}_j(z)\right), \tag{10}$$

где $C_j^+\in\mathbb{C}$, $\tilde{C}_j(z)$ – какие-либо первообразные функций $\dfrac{W_j^+(z)F_+(z)}{W_+^2(z)}$, $j=\overline{1,m}$. Для дальнейших рассуждений зафиксируем точку $z_0^+\in D_+$, $z_0^+\ne z_*$. Для определенности будем считать, что для $j=\overline{1,\gamma}$ $k_j\ge 0$, а для $j=\overline{\gamma+1,m}$ $k_j<0$. На основании порядков в точке z_* функций $W_j^+(z)$, $F_+(z)$, $W_+(z)$, установленных в лемме 1, и последующих после ее доказательства рассуждений заключаем, что порядок функции $\dfrac{W_j^+(z)F_+(z)}{W_+^2(z)}$ не ниже

$$\sum_{\substack{s=1\\ s\ne j}}^{m}k_s-\frac{(m-2)(m-1)}{2}+\sum_{j=1}^{m}k_j-\frac{(m-1)m}{2}-2\left(\sum_{j=1}^{m}k_j-\frac{m(m+1)}{2}\right)=-k_j-1.$$

Теперь понятно, что формула (10), вообще говоря, не даст аналитическую функцию в D_+ по трем причинам.

Во-первых, для $j=\overline{1,\gamma}$ упомянутые производные могут не существовать. Для их существования необходимо и достаточно выполнения условий

$$\operatorname*{res}_{z=z_*}\frac{W_j^+(z)F_+(z)}{W_+^2(z)}=0,\ j=\overline{1,\gamma}. \tag{11}$$

Если эти условия выполняются, то в качестве нужных первообразных можно брать функции $\tilde{C}_j(z)=\int_{z_0^+}^{z}\dfrac{W_j^+(\zeta)F_+(\zeta)d\zeta}{W_+^2(\zeta)}$, в которых интегралы вычисляются по любым кривым в D_+, не проходящим через точку z_*. Полюсы порядков не более k_j для $k_j>0$ у этих первообразных «погасятся» в формуле (10) нулями порядков k_j соответствующих множителей $p_j(z)$.

Во-вторых, слагаемые $C_j^+p_j(z)$ для $j=\overline{\gamma+1,m}$ в формуле (10) дадут полюс в точке z_*. Чтобы его устранить, для этих j следует брать $C_j^+=0$.

В-третьих, слагаемые $p_j(z)\tilde{C}_j(z)$ для $j=\overline{\gamma+1,\,m}$ в формуле (10) также могут дать полюс в точке z_* из-за множителей $p_j(z)$. Взяв для этих j первообразные в виде $\int\limits_{z_*}^{z}\frac{W_j^+(\zeta)F_+(\zeta)d\zeta}{W_+^2(\zeta)}$, мы получим у данных интегралов нули порядков не ниже $-k_j$ в точке z_*, «гасящие» полюсы.

Итак, для решения дифференциального уравнения (2) в классе аналитических в D_+ функций необходимо и достаточно выполнения условий (11). При их выполнении решение записывается по формуле

$$\Phi_+(z)=\sum_{j=1}^{\gamma}p_j(z)\left(C_j^+ + \int\limits_{z_0^+}^{z}\frac{W_j^+(\zeta)F_+(\zeta)d\zeta}{W_+^2(\zeta)}\right)+\sum_{j=\gamma+1}^{m}p_j(z)\int\limits_{z_*}^{z}\frac{W_j^+(\zeta)F_+(\zeta)d\zeta}{W_+^2(\zeta)}.$$

Рассуждения при решении уравнения (3) аналогичны. Для существования этого решения (с учетом условия $\Phi_-(\infty)=0$) необходимо и достаточно выполнения условий

$$\operatorname*{res}_{z=\infty}\frac{W_j^-(z)F_-(z)}{W_-^2(z)}=0,\ \ j=\overline{1,\,n}. \tag{12}$$

При их выполнении решение записывается по формуле

$$\Phi_-(z)=\sum_{j=1}^{n}q_j(z)\left(C_j^- + \int\limits_{z_0^-}^{z}\frac{W_j^-(\zeta)F_-(\zeta)d\zeta}{W_-^2(\zeta)}\right),$$

где C_j^- – произвольные постоянные, z_0^- – фиксированная точка, $z_0^-\in D_-$, $z_0^-\neq\infty$, а интегралы берутся по любым кривым в D_-, не проходящим через точку $z=\infty$.

Теперь при разных значениях $\varkappa$, α, β теореме можно придавать развернутые формулировки, которые необходимо расценивать как следствия из теоремы. В формулировках надо учитывать, что равенства (11), (12) при наличии произвольных постоянных в формулах для $F_\pm(z)$ станут линейными алгебраическими уравнениями. Например, будет справедливо нижеприведенное следствие.

Следствие. *При* $\varkappa-\beta\geq 0$, $\alpha\leq -1$ *для разрешимости уравнения (1) необходима и достаточна совместность системы*

$$\begin{cases}\sum\limits_{k=0}^{\varkappa-\beta}\alpha_{jk}c_k+\sum\limits_{k=0}^{|\alpha|-1}\beta_{jk}d_k=A_j,\ \ j=\overline{1,\,\gamma},\\ \sum\limits_{k=0}^{\varkappa-\beta}\gamma_{jk}c_k+\sum\limits_{k=0}^{|\alpha|-1}\delta_{jk}d_k=B_j,\ \ j=\overline{1,\,n},\end{cases} \tag{13}$$

где для всех указанных j, k

$$\alpha_{jk}=\operatorname*{res}_{z=z_*}\frac{W_j^+(z)X_+(z)(z-z_*)^k}{W_+^2(z)},\ \ \beta_{jk}=\operatorname*{res}_{z=z_*}\frac{W_j^+(z)X_+(z)}{W_+^2(z)(z-z_*)^{|\alpha|-k}},$$

$$A_j=-\operatorname*{res}_{z=z_*}\frac{W_j^+(z)X_+(z)\Psi_+(z)}{W_+^2(z)},\ \ \gamma_{jk}=\operatorname*{res}_{z=\infty}\frac{W_j^-(z)X_-(z)(z-z_*)^k}{W_-^2(z)},$$

$$\delta_{jk}=\operatorname*{res}_{z=\infty}\frac{W_j^-(z)X_-(z)}{W_-^2(z)(z-z_*)^{|\alpha|-k}},\ \ B_j=-\operatorname*{res}_{z=\infty}\frac{W_j^-(z)X_-(z)\Psi_-(z)}{W_-^2(z)}.$$

В случае совместности системы (13) решение уравнения записывается по формуле

$$\varphi(t)=\sum_{j=1}^{\gamma}p_j(t)\left(C_j^+ + \int\limits_{z_0^+}^{t}\frac{W_j^+(\zeta)F_+(\zeta)d\zeta}{W_+^2(\zeta)}\right)+\sum_{j=\gamma+1}^{m}p_j(t)\int\limits_{z_*}^{t}\frac{W_j^+(\zeta)F_+(\zeta)d\zeta}{W_+^2(\zeta)}-$$

$$-\sum_{j=1}^{n}q_j(t)\left(C_j^- + \int\limits_{z_0^-}^{t}\frac{W_j^-(\zeta)F_-(\zeta)d\zeta}{W_-^2(\zeta)}\right),$$

в которой постоянные c_k, d_k, входящие в выражения для $F_{\pm}(t)$, являются общим решением системы (13).

Аналогичные утверждения для иных $\varkappa$, α, β приводить не будем.

Пример. Решим уравнение

$$\begin{vmatrix} t & \sin t & \varphi(t) \\ 1 & \cos t & \varphi'(t) \\ 0 & -\sin t & \varphi''(t) \end{vmatrix} + (t+1)^{10} \begin{vmatrix} \frac{1}{t^2} & \frac{2t-1}{2t^4} & \varphi(t) \\ -\frac{2}{t^3} & \frac{2-3t}{t^5} & \varphi'(t) \\ \frac{6}{t^4} & \frac{2(6t-5)}{t^6} & \varphi''(t) \end{vmatrix} +$$

$$+ \frac{1}{\pi i} \begin{vmatrix} t & \sin t & \int\limits_{|\tau|=\frac{3}{2}} \frac{\varphi(\tau)d\tau}{\tau - t} \\ 1 & \cos t & \int\limits_{|\tau|=\frac{3}{2}} \frac{\varphi(\tau)d\tau}{(\tau - t)^2} \\ 0 & -\sin t & 2\int\limits_{|\tau|=\frac{3}{2}} \frac{\varphi(\tau)d\tau}{(\tau - t)^3} \end{vmatrix} - \frac{(t+1)^{10}}{\pi i} \begin{vmatrix} \frac{1}{t^2} & \frac{2t-1}{2t^4} & \int\limits_{|\tau|=\frac{3}{2}} \frac{\varphi(\tau)d\tau}{\tau - t} \\ -\frac{2}{t^3} & \frac{2-3t}{t^5} & \int\limits_{|\tau|=\frac{3}{2}} \frac{\varphi(\tau)d\tau}{(\tau - t)^2} \\ \frac{6}{t^4} & \frac{2(6t-5)}{t^6} & 2\int\limits_{|\tau|=\frac{3}{2}} \frac{\varphi(\tau)d\tau}{(\tau - t)^3} \end{vmatrix} = \frac{2(t+\delta)}{t-1}, \ |t| = \frac{3}{2}. \qquad (14)$$

Здесь $m = n = 2$, δ – числовой параметр, $W_+(z) = z\cos z - \sin z$, $W_-(z) = \frac{1-z}{z^7}$. Точка $z = \infty$, очевидно, будет единственной особой точкой в области $D_- = \left\{ z \middle| |z| > \frac{3}{2} \right\}$. В области $D_+ = \left\{ z \middle| |z| < \frac{3}{2} \right\}$ роль особой точки играет точка $z_* = 0$, однако единственность такой точки нуждается в обосновании.

Рассмотрим функцию $z\cos z - \sin z$ внутри квадрата с центром в нуле, стороны которого имеют длину $\frac{3\pi}{2}$ и параллельны действительной и мнимой осям. Несложно вычислить, что для всех точек на сторонах квадрата справедливо неравенство $|z\cos z| > |\sin z|$. По теореме Руше получается, что в этом квадрате число нулей функции $z\cos z$ равно числу нулей функции $z\cos z - \sin z$. Функция $z\cos z$, очевидно, имеет нули лишь в точках $0, \pm\frac{\pi}{2}$ квадрата, а в область D_+ попадает только точка $z_* = 0$. (В связи с нулями функции $z\cos z - \sin z$ см. также пример 23.17 в сборнике задач [11].)

Краевое условие задачи Римана (4) примет вид

$$\begin{vmatrix} t - \sin t & \sin t & \Phi_+(t) \\ 1 - \cos t & \cos t & \Phi'_+(t) \\ \sin t & -\sin t & \Phi''_+(t) \end{vmatrix} = (t+1)^{10} \begin{vmatrix} \frac{1}{t^2} & \frac{2t-1}{2t^4} & \Phi_-(t) \\ -\frac{2}{t^3} & \frac{2-3t}{t^5} & \Phi'_-(t) \\ \frac{6}{t^4} & \frac{2(6t-5)}{t^6} & \Phi''_-(t) \end{vmatrix} + \frac{t+\delta}{t-1}, \ |t| = \frac{3}{2}. \qquad (15)$$

При этом в определителе, стоящем в левой части (15), исходный первый столбец заменен на разность первого и второго столбцов. Тем самым в согласии с общей схемой созданы разные порядки в точке $z_* = 0$ у первых элементов этих столбцов. Вычисления показывают, что $m_1 = 3$, $m_2 = 1$, $\alpha = 1$, $n_1 = 2$, $n_2 = 3$, $\beta = 9$. Следовательно, краевую задачу Римана

$$F_+(t) = (t+1)^{10} F_-(t) + \frac{t+\delta}{t-1}, \ |t| = \frac{3}{2},$$

для функций

$$F_+(z)=\begin{vmatrix} z-\sin z & \sin z & \Phi_+(z)\\ 1-\cos z & \cos z & \Phi_+'(z)\\ \sin z & -\sin z & \Phi_+''(z)\end{vmatrix},\ F_-(z)=\begin{vmatrix} \frac{1}{z^2} & \frac{2z-1}{2z^4} & \Phi_-(z)\\ -\frac{2}{z^3} & \frac{2-3z}{z^5} & \Phi_-'(z)\\ \frac{6}{z^4} & \frac{2(6z-5)}{z^6} & \Phi_-''(z)\end{vmatrix}$$

необходимо решать в классе функций, кратных дивизору $0^1\infty^9$. В результате получаем

$$F_+(z)=cz,\ F_-(z)=\frac{cz^2-(c+1)z-\delta}{(z+1)^{10}(z-1)},\ \forall c\in\mathbb{C}.$$

Теперь следует решать в D_+ дифференциальное уравнение

$$\begin{vmatrix} z-\sin z & \sin z & \Phi_+(z)\\ 1-\cos z & \cos z & \Phi_+'(z)\\ \sin z & -\sin z & \Phi_+''(z)\end{vmatrix}=cz,$$

откуда находим

$$\Phi_+(z)=C_1^+(z-\sin z)+C_2^+\sin z+$$
$$+c(\sin z-z)\int\limits_{z_0^+}^{z}\frac{\zeta\sin\zeta\,d\zeta}{(\zeta\cos\zeta-\sin\zeta)^2}+c\sin z\int\limits_{z_0^+}^{z}\frac{(\zeta-\sin\zeta)\zeta\,d\zeta}{(\zeta\cos\zeta-\sin\zeta)^2}.$$

Обе подынтегральные функции в последней формуле являются четными, поэтому их вычеты в точке $z_* = 0$ равны нулю, так что интегралы будут давать однозначные функции. Далее следует решать в D_- уравнение

$$\begin{vmatrix} \frac{1}{z^2} & \frac{2z-1}{2z^4} & \Phi_-(z)\\ -\frac{2}{z^3} & \frac{2-3z}{z^5} & \Phi_-'(z)\\ \frac{6}{z^4} & \frac{2(6z-5)}{z^6} & \Phi_-''(z)\end{vmatrix}=\frac{cz^2-(c+1)z-\delta}{(z+1)^{10}(z-1)}.$$

Решением будет функция

$$\Phi_-(z)=\frac{C_1^-}{z^2}+\frac{C_2^-(2z-1)}{2z^4}+\frac{1}{2z^2}\int\limits_{z_0^-}^{z}\frac{(2\zeta-1)\left(c\zeta^2-(c+1)\zeta-\delta\right)\zeta^{10}d\zeta}{(\zeta+1)^{10}(1-\zeta)^3}+$$
$$+\frac{2z-1}{2z^4}\int\limits_{z_0^-}^{z}\frac{\left(c\zeta^2-(c+1)\zeta-\delta\right)\zeta^{12}d\zeta}{(\zeta+1)^{10}(\zeta-1)^3},$$

где постоянные c, δ еще нужно подобрать так, чтобы указанные интегралы давали однозначные функции в D_-. Для этого необходимо и достаточно выполнения условий

$$\operatorname*{res}_{z=\infty}\frac{(2z-1)\left(cz^2-(c+1)z-\delta\right)z^{10}}{(z+1)^{10}(1-z)^3}=-17c-2=0,$$

$$\operatorname*{res}_{z=\infty}\frac{\left(cz^2-(c+1)z-\delta\right)z^{12}}{(z+1)^{10}(z-1)^3}=\delta-38c-7=0,$$

откуда $c=-\frac{2}{17}$, $\delta=\frac{43}{17}$. Итак, уравнение (14) разрешимо лишь при $\delta=\frac{43}{17}$. Решение при этом значении δ, найденное по формуле (5) с учетом получившегося значения c, будет равно

$$\varphi(t) = C_1^+(t - \sin t) + C_2^+ \sin t - \frac{C_1^-}{t^2} - \frac{C_2^-(2t-1)}{2t^4} +$$

$$+ \frac{2}{17}(t - \sin t)\int_{z_0^+}^{t} \frac{\zeta \sin\zeta d\zeta}{(\zeta\cos\zeta - \sin\zeta)^2} - \frac{2\sin t}{17}\int_{z_0^+}^{t} \frac{(\zeta - \sin\zeta)\zeta\, d\zeta}{(\zeta\cos\zeta - \sin\zeta)^2} +$$

$$+ \frac{1}{34t^2}\int_{z_0^-}^{t} \frac{(2\zeta - 1)\left(2\zeta^2 + 15\zeta + 43\right)\zeta^{10} d\zeta}{(\zeta+1)^{10}(1-\zeta)^3} - \frac{2t-1}{34t^4}\int_{z_0^-}^{t} \frac{\left(2\zeta^2 + 15\zeta + 43\right)\zeta^{12} d\zeta}{(\zeta+1)^{10}(1-\zeta)^3},\ |t| = \frac{3}{2}. \qquad (16)$$

Отметим, что два последних интеграла в (16) поддаются дальнейшим вычислениям. Результаты этих вычислений, полученные с помощью подходящих компьютерных программ, приводить в настоящей статье представляется нецелесообразным из-за их громоздкости.

Заключение

Исследование исходного уравнения носит законченный характер. Дальнейшие разработки возможны для случая нескольких особых точек. По-видимому, можно вводить в рассмотрение и конструктивно исследовать уравнения с определителями, сводящиеся к краевым задачам Гильберта, Карлемана и др.

Библиографические ссылки

1. Boykov IV, Ventsel ES, Boykova AI. An approximate solution of hypersingular integral equations. *Applied Numerical Mathematics.* 2010;60(6):607–628. DOI: 10.1016/j.apnum.2010.03.003.

2. Бойков ИВ. Аналитические и численные методы решения гиперсингулярных интегральных уравнений. *Динамические системы.* 2019;9(3):244–272.

3. Зверович ЭИ. Решение гиперсингулярного интегро-дифференциального уравнения с постоянными коэффициентами. *Доклады Национальной академии наук Беларуси.* 2010;54(6):5–8.

4. Зверович ЭИ, Шилин АП. Решение интегро-дифференциальных уравнений с сингулярными и гиперсингулярными интегралами специального вида. *Известия Национальной академии наук Беларуси. Серия физико-математических наук.* 2018; 54(4):404–407. DOI: 10.29235/1561-2430-2018-54-4-404-407.

5. Шилин АП. Гиперсингулярное интегро-дифференциальное уравнение эйлерова типа. *Известия Национальной академии наук Беларуси. Серия физико-математических наук.* 2020;56(1):17–29. DOI: 10.29235/1561-2430-2020-56-1-17-29.

6. Шилин АП. О решении одного интегро-дифференциального уравнения с сингулярным и гиперсингулярным интегралами. *Известия Национальной академии наук Беларуси. Серия физико-математических наук.* 2020;56(3):298–309. DOI: 10.29235/1561-2430-2020-56-3-298-309.

7. Шилин АП. Дифференциальная краевая задача Римана и ее приложение к интегро-дифференциальным уравнениям. *Доклады Национальной академии наук Беларуси.* 2019;63(4):391–397. DOI: 10.29235/1561-8323-2019-63-4-391-397.

8. Зверович ЭИ. Краевые задачи теории аналитических функций в гёльдеровских классах на римановых поверхностях. *Успехи математических наук.* 1971;26(1):113–179.

9. Зверович ЭИ. Обобщение формул Сохоцкого. *Известия Национальной академии наук Беларуси. Серия физико-математических наук.* 2012;2:24–28.

10. Гахов ФД. *Краевые задачи.* 3-е издание. Москва: Наука; 1977. 640 с.

11. Евграфов МА, Бежанов КА, Сидоров ЮВ, Федорюк МВ, Шабунин МИ. *Сборник задач по теории аналитических функций.* 2-е издание. Евграфов МА, редактор. Москва: Наука; 1972. 416 с.

References

1. Boykov IV, Ventsel ES, Boykova AI. An approximate solution of hypersingular integral equations. *Applied Numerical Mathematics.* 2010;60(6):607–628. DOI: 10.1016/j.apnum.2010.03.003.

2. Boykov IV. Analytical and numerical methods for solving hypersingular integral equations. *Dinamicheskie sistemy.* 2019;9(3): 244–272. Russian.

3. Zverovich EI. Solution of the hypersingular integro-differential equation with constant coefficients. *Doklady of the National Academy of Sciences of Belarus.* 2010;54(6):5–8. Russian.

4. Zverovich EI, Shilin AP. Integro-differential equations with singular and hypersingular integrals. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series.* 2018;54(4):404–407. Russian. DOI: 10.29235/1561-2430-2018-54-4-404-407.

5. Shilin AP. A hypersingular integro-differential equation of the Euler type. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series.* 2020;56(1):17–29. Russian. DOI: 10.29235/1561-2430-2020-56-1-17-29.

6. Shilin AP. On the solution of one integro-differential equation with singular and hypersingular integrals. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series.* 2020;56(3):298–309. Russian. DOI: 10.29235/1561-2430-2020-56-3-298-309.

7. Shilin AP. Riemann's differential boundary-value problem and its application to integro-differential equations. *Doklady of the National Academy of Sciences of Belarus.* 2019;63(4):391–397. Russian. DOI: 10.29235/1561-8323-2019-63-4-391-397.

8. Zverovich EI. [Boundary value problem in the theory of analytic functions in Holder classes on Riemann surfaces]. *Uspekhi matematicheskikh nauk.* 1971;26(1):113–179. Russian.

9. Zverovich EI. Generalization of Sokhotski's formulas. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series.* 2012;2:24–28. Russian.

10. Gakhov FD. *Kraevye zadachi* [Boundary value problems]. 3rd edition. Moscow: Nauka; 1977. 640 p. Russian.

11. Yevgrafov MA, Bezhanov KA, Sidorov YuV, Fedoryuk MV, Shabunin MI. *Sbornik zadach po teorii analiticheskikh funktsii* [Collection of problems on the theory of analytic functions]. 2nd edition. Yevgrafov MA, editor. Moscow: Nauka; 1972. 416 p. Russian.

Получена 04.05.2021 / исправлена 02.06.2021 / принята 02.06.2021.
Received 04.05.2021 / revised 02.06.2021 / accepted 02.06.2021.

УДК 517.547.59

О СВОЙСТВАХ *h*-ДИФФЕРЕНЦИРУЕМЫХ ФУНКЦИЙ

В. А. ПАВЛОВСКИЙ [1)]***, И. Л. ВАСИЛЬЕВ*** [1)]

[1)] *Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Исследования в области теории функций *h*-комплексной переменной представляют интерес в связи с имеющимися приложениями в неевклидовой геометрии, теоретической механике и т. д. Изучены свойства *h*-дифференцируемых функций. Найдены критерии *h*-дифференцируемости и *h*-голоморфности, сформулирована и доказана теорема о конечных приращениях для *h*-голоморфной функции. Приведены достаточные условия *h*-аналитичности, сформулирована и доказана теорема единственности для *h*-аналитических функций.

Ключевые слова: кольцо *h*-комплексных чисел; делители нуля; *h*-дифференцируемость; *h*-голоморфность; *h*-аналитичность; конечные приращения функции; нули функции; ряд Тейлора.

ON PROPERTIES OF *h*-DIFFERENTIABLE FUNCTIONS

V. A. PAVLOVSKY [a]***, I. L. VASILIEV*** [a]

[a] *Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

Corresponding author: V. A. Pavlovsky (pavlad95@gmail.com)

Research in the theory of functions of an *h*-complex variable is of interest in connection with existing applications in non-Euclidean geometry, theoretical mechanics, etc. This article is devoted to the study of the properties of *h*-differentiable functions. Criteria for *h*-differentiability and *h*-holomorphy are found, formulated and proved a theorem on finite increments for an *h*-holomorphic function. Sufficient conditions for *h*-analyticity are given, formulated and proved a uniqueness theorem for *h*-analytic functions.

Keywords: ring of *h*-complex numbers; zero divisors; *h*-differentiability; *h*-holomorphy; *h*-analyticity; finite increments of a function; zeros of a function; Taylor series.

Множество *h*-комплексных чисел

Пусть $\mathbb{C}_h$ – множество всех *h*-комплексных (двойных) чисел [1–7], т. е. множество упорядоченных пар вещественных чисел, на котором заданы операции сложения и умножения $\forall z_1 = (a; b),\ z_2 = (c; d) \in \mathbb{C}_h$ по правилам:

1) $z_1 + z_2 = (a + c; b + d)$;

2) $z_1 \cdot z_2 = (ac + bd; ad + bc)$.

Образец цитирования:
Павловский ВА, Васильев ИЛ. О свойствах *h*-дифференцируемых функций. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:29–37. https://doi.org/10.33581/2520-6508-2021-2-29-37

For citation:
Pavlovsky VA, Vasiliev IL. On properties of *h*-differentiable functions. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:29–37. Russian. https://doi.org/10.33581/2520-6508-2021-2-29-37

Авторы:
Владислав Андреевич Павловский – аспирант кафедры теории функций механико-математического факультета. Научный руководитель – И. Л. Васильев.
Игорь Леонидович Васильев – кандидат физико-математических наук, доцент; доцент кафедры теории функций механико-математического факультета.

Authors:
Vladislav A. Pavlovsky, postgraduate student at the department of function theory, faculty of mechanics and mathematics.
pavlad95@gmail.com
Igor L. Vasiliev, PhD (physics and mathematics), docent; associate professor at the department of function theory, faculty of mechanics and mathematics.
vassilyevi@bsu.by

Вещественную единицу отождествим с h-комплексным числом $(1; 0)$. Гиперболической единицей назовем h-комплексное число $j = (0; 1)$. Тогда любое число из $\mathbb{C}_h$ может быть представлено в алгебраической форме:

$$z = (a; b) = (a; 0) + (0; b) = a \cdot (1; 0) + b \cdot (0; 1) = a + jb = \mathrm{Re}\, z + j\,\mathrm{Hyp}\, z,$$

где $a = \mathrm{Re}\, z$ – вещественная часть числа z, а $b = \mathrm{Hyp}\, z$ – гиперболическая часть числа z.

Как показано в работе [6], множество h-комплексных чисел $\mathbb{C}_h$ есть кольцо с делителями нуля, каковыми являются числа вида $a \pm aj$. Особо стоит отметить случай, когда $a = \frac{1}{2}$. Тогда делители нуля обладают следующими свойствами:

- $\left(\frac{1 \pm j}{2}\right)^n = \frac{1 \pm j}{2}\ \forall n \in \mathbb{N}$;
- числа $\frac{1 \pm j}{2}$ образуют базис в $\mathbb{C}_h$, т. е. любое h-комплексное число $a + jb$ можно однозначно представить в виде

$$a + jb = (a + b)\frac{1 + j}{2} + (a - b)\frac{1 - j}{2}.$$

Норму элемента $z = a + jb$ в кольце $\mathbb{C}_h$ определим следующим образом: $\|z\| = |a| + |b|$, а модулем h-комплексного числа назовем, как обычно, $|z| = \sqrt{a^2 + b^2}$.

Приведем свойства нормы:

1) $\|z\| = 0 \Leftrightarrow z = 0$;

2) $|z| = \sqrt{a^2 + b^2} \le |a| + |b| = \|z\| \le \sqrt{2}\sqrt{a^2 + b^2} = \sqrt{2}\,|z|$;

3) $\|\alpha z\| = |\alpha| \cdot \|z\|\ \forall \alpha \in \mathbb{R}$;

4) $\|z_1 \cdot z_2\| \le \|z_1\| \cdot \|z_2\|\ \forall z_1,\ z_2 \in \mathbb{C}_h$;

5) $\|z^n\| = \|z\|^n\ \forall n \in \mathbb{N}$;

6) $\frac{1}{\|z\|} \le \left\|\frac{1}{z}\right\|$.

На множестве $\mathbb{C}_h$ топология вводится с помощью вышеуказанной нормы.

h-Дифференцируемость функций

Пусть D – область в $\mathbb{C}_h$, а $f : D \to \mathbb{C}_h$.

Определение 1. Функция f называется h-дифференцируемой в точке $z \in D$, если существует такое число $k \in \mathbb{C}_h$, что

$$f(z + h) - f(z) = kh + \alpha(h)h, \tag{1}$$

где $h \in D$ не является делителем нуля, причем $z + h \in D$, а $\lim\limits_{h \to 0} \alpha(h) = 0$, k не зависит от h.

Определение 2. Производной функции f h-комплексного аргумента $z \in D$ называется выражение вида

$$f'(z) = \lim_{h \to 0} \frac{f(z + h) - f(z)}{h}, \tag{2}$$

где $h \in \mathbb{C}_h$ не является делителем нуля. Предел берется по норме из $\mathbb{C}_h$.

Производные суммы, разности, произведения, частного от деления и композиции функций вычисляются по тем же формулам, что и в классическом анализе, при условии, что они определены.

Теорема 1. *Функция $f(z)$ h-дифференцируема в точке $z \in D$ тогда и только тогда, когда выполняется (2).*

Доказательство проводится так же, как и в случае аналитической функции комплексного переменного, при этом $f'(z) = k$ из (1).

Любая h-комплексная функция $f(z) = f(x + jy)$ представима в алгебраической форме:

$$f(z) = u(x, y) + jv(x, y).$$

Теорема 2. *Пусть функция* $f(z)=u(x,y)+jv(x,y)$ *определена в окрестности точки* $z=x+jy$, *функции* $u(x,y)$ *и* $v(x,y)$ *дифференцируемы в точке* (x,y). *Тогда следующие утверждения эквивалентны:*

1) функция $f(z)$ *h-дифференцируема в точке z;*

2) в точке (x,y) *верны равенства*

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y},\ \frac{\partial u}{\partial y}=\frac{\partial v}{\partial x}. \tag{3}$$

Доказательство. Покажем, что из первого утверждения следует второе. Пусть $h=s+jt$,

$$f'(z)=\lim_{h\to 0}\frac{f(z+h)-f(z)}{h}.$$

Положим $t=0$. В этом случае

$$f'(z)=\lim_{s\to 0}\frac{u(x+s,y)-u(x,y)}{s}+j\lim_{s\to 0}\frac{v(x+s,y)-v(x,y)}{s}=\frac{\partial u}{\partial x}+j\frac{\partial v}{\partial x}.$$

Пусть $s=0$, тогда

$$f'(z)=\lim_{t\to 0}\frac{u(x,y+t)-u(x,y)}{jt}+j\lim_{t\to 0}\frac{v(x,y+t)-v(x,y)}{jt}=\frac{\partial v}{\partial y}+j\frac{\partial u}{\partial y}.$$

Имеем

$$\frac{\partial u}{\partial x}+j\frac{\partial v}{\partial x}=\frac{\partial v}{\partial y}+j\frac{\partial u}{\partial y},$$

следовательно, верны равенства (3).

Теперь покажем, что из второго утверждения следует первое. Пусть верны равенства (3), тогда

$$f(z+h)-f(z)=\left[u(x+s,y+t)-u(x,y)\right]+j\left[v(x+s,y+t)-v(x,y)\right]=$$

$$=\left(u'_x s+u'_y t+\alpha(h)h\right)+j\left(v'_x s+v'_y t+\beta(h)h\right)=u'_x(s+jt)+$$

$$+jv'_x(s+jt)+\left(\alpha(h)+j\beta(h)\right)h=\left(u'_x+jv'_x\right)h+\gamma(h)h,$$

где $\gamma(h)=\alpha(h)+j\beta(h)$, $\lim\limits_{h\to 0}\gamma(h)=0$, следовательно, функция $f(z)$ *h*-дифференцируема,

$$f'(z)=u'_x+jv'_x=v'_y+ju'_y.$$

Теорема доказана.

Замечание. Равенства (3) являются аналогом условий Коши – Римана.

h-Голоморфность функций

Определение 3. Функция $f(z)=u(x,y)+jv(x,y)$ называется *h*-голоморфной в точке $z_0=x_0+jy_0\in D$, если в некоторой окрестности этой точки функции *u* и *v* имеют непрерывные вторые частные производные и выполнены условия (3).

Введем обозначение $D^*=\left\{(x,y)\in\mathbb{R}^2\,|\,z=x+jy\in D\right\}$ и далее будем считать, что функции *u* и *v* дважды непрерывно дифференцируемы в D^*.

Теорема 3. *Функция f является h-голоморфной в точке* $z\in D$ *тогда и только тогда, когда*

$$f(z)=\frac{1+j}{2}f(x+y)+\frac{1-j}{2}f(x-y). \tag{4}$$

Доказательство. Рассмотрим функцию $f(z)=u(x,y)+jv(x,y)$. Пусть выполнены условия (3), тогда функции *u* и *v* удовлетворяют уравнениям

$$\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}=0,\ \frac{\partial^2 v}{\partial x^2}-\frac{\partial^2 v}{\partial y^2}=0. \tag{5}$$

Положим $\xi = \frac{1}{2}(x+y)$, $\eta = \frac{1}{2}(x-y)$. В этом случае

$$\begin{cases} \dfrac{\partial u}{\partial \xi} = \dfrac{\partial u}{\partial x} x'_\xi + \dfrac{\partial u}{\partial y} y'_\xi = u'_x + u'_y, \\ \dfrac{\partial u}{\partial \eta} = \dfrac{\partial u}{\partial x} x'_\eta + \dfrac{\partial u}{\partial y} y'_\eta = u'_x - u'_y. \end{cases}$$

Смешанные производные функций u и v равны нулю:

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0, \ \frac{\partial^2 u}{\partial \eta \partial \xi} = 0.$$

Таким образом, уравнения (5) равносильны следующим уравнениям:

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = \frac{\partial^2 u}{\partial \eta \partial \xi} = 0. \tag{5а}$$

Аналогично получаем уравнения для функции v:

$$\frac{\partial^2 v}{\partial \xi \partial \eta} = \frac{\partial^2 v}{\partial \eta \partial \xi} = 0. \tag{5б}$$

Найдем общее решение уравнений (5а) и (5б):

$$u'_\xi = \mu^*(\xi),$$

$$u(\xi, \eta) = \int \mu^*(\xi) d\xi = \tilde{\mu}(\xi) + \tilde{\psi}(\eta) =$$

$$= \tilde{\mu}\left(\frac{x+y}{2}\right) + \tilde{\psi}\left(\frac{x-y}{2}\right) = \frac{1}{2}\{\mu(x+y) + \psi(x-y)\},$$

$$v'_\xi = \varphi^*(\xi),$$

$$v(\xi, \eta) = \int \varphi^*(\xi) d\xi = \tilde{\varphi}(\xi) + \tilde{\nu}(\eta) =$$

$$= \tilde{\varphi}\left(\frac{x+y}{2}\right) + \tilde{\nu}\left(\frac{x-y}{2}\right) = \frac{1}{2}\{\varphi(x+y) + \nu(x-y)\}.$$

Из уравнений

$$u'_x = v'_y, \ v'_x = u'_y$$

вытекает

$$\begin{cases} \frac{1}{2}\{\mu'(x+y) + \psi'(x-y)\} = \frac{1}{2}\{\varphi'(x+y) - \nu'(x-y)\}, \\ \frac{1}{2}\{\mu'(x+y) - \psi'(x-y)\} = \frac{1}{2}\{\varphi'(x+y) + \nu'(x-y)\}, \end{cases}$$

следовательно,

$$\begin{cases} \mu'(x+y) = \varphi'(x+y), \\ \psi'(x-y) = \nu'(x-y), \end{cases} \quad \begin{cases} \mu(x+y) = \varphi(x+y) + \alpha, \\ \psi(x-y) = \nu(x-y) + \beta, \end{cases}$$

$$\begin{cases} u(x, y) = \frac{1}{2}\{\varphi(x+y) + \psi(x-y) + \alpha\}, \\ v(x, y) = \frac{1}{2}\{\varphi(x+y) - \psi(x-y) + \beta\}. \end{cases}$$

Имеем

$$\begin{cases} f(z) = f(x+jy) = u(x, y) + jv(x, y), \\ \overline{f}(z) = \overline{f}(x+jy) = u(x, y) - jv(x, y), \end{cases} \quad \begin{cases} f(x) = u(x, 0) + jv(x, 0), \\ \overline{f}(x) = u(x, 0) - jv(x, 0), \end{cases}$$

из этого следует, что

$$\begin{cases} u(x,0)=\frac{1}{2}\{f(x)+\overline{f}(x)\}, \\ v(x,0)=\frac{1}{2}\{f(x)-\overline{f}(x)\}, \end{cases} \quad \begin{cases} u(x,0)=\frac{1}{2}\{\varphi(x)+\psi(x)+\alpha\}, \\ v(x,0)=\frac{1}{2}\{\varphi(x)-\psi(x)+\beta\}, \end{cases}$$

а значит,

$$\begin{cases} \varphi(x)=u(x,0)+v(x,0)-\frac{\alpha+\beta}{2}, \\ \psi(x)=u(x,0)-v(x,0)-\frac{\alpha-\beta}{2}, \end{cases}$$

тогда

$$\begin{cases} \varphi(x)=\frac{1}{2}\left(f(x)+\overline{f}(x)\right)+\frac{j}{2}\left(f(x)-\overline{f}(x)\right)-\frac{\alpha+\beta}{2}, \\ \psi(x)=\frac{1}{2}\left(f(x)+\overline{f}(x)\right)-\frac{j}{2}\left(f(x)-\overline{f}(x)\right)-\frac{\alpha-\beta}{2}, \end{cases}$$

$$\begin{cases} \varphi(x)=\frac{1+j}{2}f(x)+\frac{1-j}{2}\overline{f}(x)-\frac{\alpha+\beta}{2}, \\ \psi(x)=\frac{1-j}{2}f(x)+\frac{1+j}{2}\overline{f}(x)-\frac{\alpha-\beta}{2}. \end{cases}$$

Отсюда находим, что

$$f(z)=u(x,y)+jv(x,y)=\frac{1}{2}\{\varphi(x+y)+\psi(x-y)+\alpha\}+\frac{j}{2}\{\varphi(x+y)-\psi(x-y)+\beta\},$$

$$f(z)=\frac{1}{2}\left\{\frac{1+j}{2}f(x+y)+\frac{1-j}{2}\overline{f}(x+y)-\frac{\alpha+\beta}{2}+\alpha+\frac{1+j}{2}f(x-y)+\frac{1-j}{2}\overline{f}(x-y)-\frac{\alpha-\beta}{2}\right\}+$$

$$+\frac{j}{2}\left\{\frac{1+j}{2}f(x+y)+\frac{1-j}{2}\overline{f}(x+y)-\frac{\alpha+\beta}{2}-\frac{1+j}{2}f(x-y)-\frac{1-j}{2}\overline{f}(x-y)+\frac{\alpha-\beta}{2}+\beta\right\}=$$

$$=\frac{1+j}{4}f(x+y)+\frac{1-j}{4}\overline{f}(x+y)+\frac{1-j}{4}f(x-y)+\frac{1+j}{4}\overline{f}(x-y)+\frac{1+j}{4}f(x+y)-$$

$$-\frac{1-j}{4}\overline{f}(x+y)+\frac{1-j}{4}f(x-y)-\frac{1+j}{4}\overline{f}(x-y)=\frac{1+j}{2}f(x+y)+\frac{1-j}{2}f(x-y).$$

Таким образом, справедливо равенство (4).

Обратно, пусть верно (4), тогда для функции $f(z)=u(x,y)+jv(x,y)$ положим $y=0$:

$$f(x)=u(x,0)+jv(x,0),$$

тогда

$$\begin{cases} f(x+y)=u(x+y,0)+jv(x+y,0), \\ f(x-y)=u(x-y,0)+jv(x-y,0). \end{cases}$$

Используя равенство (4), представим функцию $f(z)$ в виде

$$f(z)=\frac{1+j}{2}\left[u(x+y,0)+jv(x+y,0)\right]+\frac{1-j}{2}\left[u(x-y,0)+jv(x-y,0)\right]=$$

$$=\frac{1}{2}\left[u(x+y,0)+v(x+y,0)+u(x-y,0)-v(x-y,0)\right]+$$

$$+\frac{j}{2}\left[u(x+y,0)+v(x+y,0)-u(x-y,0)+v(x-y,0)\right]=u(x,y)+jv(x,y).$$

В силу

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y},\quad \frac{\partial u}{\partial y}=\frac{\partial v}{\partial x}$$

получаем, что теорема доказана.

Теорема 4. *Пусть функция f h-голоморфна в области $D\subset\mathbb{C}_h$ с кусочно-гладким краем ∂D и непрерывна в замыкании $\bar{D}=D\cup\partial D$. Тогда*

$$\int\limits_{\partial D} f(z)\,dz=0.$$

Доказательство.

$$\int\limits_{\partial D} f(z)\,dz=\int\limits_{\partial D}\big[u(x,y)+jv(x,y)\big](dx+jdy)=$$

$$=\int\limits_{\partial D} u(x,y)\,dx+v(x,y)\,dy+j\int\limits_{\partial D} u(x,y)\,dy+v(x,y)\,dx,$$

используя формулу Грина, получаем

$$\int\limits_{\partial D} f(z)\,dz=\iint\limits_{D}\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)dxdy+j\iint\limits_{D}\left(\frac{\partial u}{\partial x}-\frac{\partial v}{\partial y}\right)dxdy=0.$$

Теорема доказана.

Далее нам понадобится следующая теорема вещественного анализа, которая выводится из второй теоремы о конечных приращениях [8].

Теорема 5 (о конечных приращениях для отображений из $\mathbb{R}^2$ в $\mathbb{R}^2$). *Пусть функция $F:\tilde{D}\subset\mathbb{R}^2\to\mathbb{R}^2$ h-дифференцируема в точке $(a,b)\in\tilde{D}$. Тогда*

$$\big|F(a+s,\,b+t)-F(a,\,b)\big|\le\max_{\xi\in[0,1]}\left|F'(a+\xi s,\,b+\xi t)\begin{bmatrix}s\\t\end{bmatrix}\right|. \tag{6}$$

Доказательство. Введем вспомогательную функцию

$$g(\tau)=F(a+\tau s,\,b+\tau t),\ \ \tau\in[0,1].$$

Имеем

$$g:[0,1]\to\mathbb{R}^2,\ g(0)=F(a,\,b),\ g(1)=F(a+s,\,b+t),$$

$$g'=F'(a+\tau s,\,b+\tau t)\begin{bmatrix}s\\t\end{bmatrix}.$$

Положим

$$G(\tau)=\big\langle g(t)\big|g(1)-g(0)\big\rangle,$$

$$G:[0,1]\to\mathbb{R},\ G'(\tau)=\big\langle g'(t)\big|g(1)-g(0)\big\rangle.$$

Отсюда по теореме Лагранжа следует, что

$$G(1)-G(0)=G'(\xi)\cdot 1,\ \text{где}\ \xi\in[0,1].$$

Используя неравенство Коши для скалярного произведения, получаем

$$G(1)-G(0)=\big\langle g(1)\big|g(1)-g(0)\big\rangle-\big\langle g(0)\big|g(1)-g(0)\big\rangle=\big\langle g(1)-g(0)\big|g(1)-g(0)\big\rangle=$$

$$=\big|g(1)-g(0)\big|^2=\big\langle g'(\xi)\big|g(1)-g(0)\big\rangle\le\big\langle\big|g'(\xi)\big|\big|g(1)-g(0)\big\rangle.$$

Следовательно,

$$\big|g(1)-g(0)\big|\le\max_{\xi\in[0,1]}\big|g'(\xi)\big|.$$

Это неравенство эквивалентно неравенству (6). Теорема доказана.

Представим функцию $F(x,\,y)$ в векторном виде: $F(x,\,y)=\begin{bmatrix}u(x,\,y)\\v(x,\,y)\end{bmatrix}$, тогда

$$F'(x,\,y)=\begin{bmatrix}u'_x(x,\,y) & u'_y(x,\,y)\\ v'_x(x,\,y) & v'_y(x,\,y)\end{bmatrix}.$$

Теперь из неравенства (6) и неравенства Коши – Буняковского получаем

$$\left|F(a+s,\,b+t)-F(a,\,b)\right|=\left\|\begin{bmatrix}u(a+s,\,b+t)-u(a,\,b)\\ v(a+s,\,b+t)-v(a,\,b)\end{bmatrix}\right\|=\left\|\begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right\|=\sqrt{|\Delta u|^2+|\Delta v|^2}\le$$

$$\le\max_{\xi\in[0,1]}\sqrt{\left(u'_x s+u'_y t\right)^2+\left(v'_x s+v'_y t\right)^2}\le\max_{\xi\in[0,1]}\sqrt{\left(\left|u'_x\right|^2+\left|u'_y\right|^2+\left|v'_x\right|^2+\left|v'_y\right|^2\right)\left(s^2+t^2\right)},\qquad(7)$$

где все частные производные вычислены в точке $(a+\xi s,\,b+\xi t)$.

Пусть $f(z)=u(x,\,y)+jv(x,\,y)$ – h-голоморфная функция, тогда $u'_x=v'_y$, $u'_y=v'_x$, следовательно,

$$f'(z)=u'_x+jv'_x=u'_x+ju'_y=v'_y+jv'_x=v'_y+ju'_y,$$

$$\left|f'(z)\right|=\sqrt{\left|u'_x\right|^2+\left|v'_x\right|^2}\le\left|u'_x\right|+\left|u'_y\right|=\left\|f'(z)\right\|,$$

$$\left\|f(z+h)-f(z)\right\|=\left\|\Delta u+j\Delta v\right\|=|\Delta u|+|\Delta v|\le\sqrt{2}\sqrt{|\Delta u|^2+|\Delta v|^2},\qquad(8)$$

где $|h|=|s+jt|=\sqrt{s^2+t^2}\le|s|+|t|=\|h\|$.

Теорема 6 (о конечных приращениях для h-голоморфной функции). *Пусть функция f h-голоморфна в области $D\subset\mathbb{C}_h$. Тогда*

$$\left\|f(z+h)-f(z)\right\|\le 2\max_{\zeta\in[z,\,z+h]}\left\|f'(\zeta)\right\|\cdot\|h\|.\qquad(9)$$

Доказательство. В силу неравенств (7) и (8) имеем

$$\left\|f(z+h)-f(z)\right\|\le\sqrt{2}\sqrt{|\Delta u|^2+|\Delta v|^2}\le$$

$$\le\max_{\xi\in[0,1]}\sqrt{2\left\{\left|u'_x\right|^2+\left|u'_y\right|^2\right\}\left\{s^2+t^2\right\}}\le 2\max_{\zeta\in[z,\,z+h]}\left|f'(\zeta)\right|\cdot|h|\le 2\max_{\zeta\in[z,\,z+h]}\left\|f'(\zeta)\right\|\cdot\|h\|.$$

Теорема доказана.

h-Аналитичность функций

Определение 4. Функция f называется h-аналитической в точке $z_0\in D$, если существует некоторая окрестность этой точки, в которой f разлагается в сходящийся степенной ряд:

$$f(z)=\sum_{k=0}^{\infty}c_k(z-z_0)^k,\ c_k\in\mathbb{C}_h.\qquad(10)$$

Из определения вытекает, что h-аналитическая функция f бесконечно h-дифференцируема в некоторой окрестности точки z_0, а ряд (10) есть ряд Тейлора функции f, т. е. $c_k=\dfrac{f^{(k)}(z_0)}{k!}$.

Областью сходимости ряда (10) является открытый h-круг

$$G=\left\{\|z-z_0\|<r\right\},$$

где $r=\dfrac{1}{\overline{\lim\limits_{k\to\infty}}\sqrt[k]{\|c_k\|}}$.

Теорема 7. *Пусть функция $f: D\to\mathbb{C}_h$ является бесконечное число раз h-дифференцируемой в области $D\subset\mathbb{C}_h$,*

$$\left\|f^{(n)}(z)\right\|\le Me^{AR^m}\ \forall n\in\mathbb{N},\ \forall z\in\left\{\|z-z_0\|\le R\right\}\subset D,\qquad(11)$$

M, A, m – некоторые положительные константы. Тогда f разлагается в ряд Тейлора:

$$f(z)=\sum_{k=0}^{\infty}\frac{f^{(k)}(z_0)}{k!}(z-z_0)^k,\ z_0\in D,$$

равномерно сходящийся в круге $\|z-z_0\|\le R$.

Доказательство. Представим $f(z)$ в виде

$$f(z) = T_n(z, z_0) + r_n(z),$$

где $T_n(z, z_0) = \sum_{k=0}^{n} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k$; $r_n(z)$ – остаточный член. Составим вспомогательную функцию

$$F(t) = f(z) - T_n(z, t) = f(z) - \sum_{k=0}^{n} \frac{f^{(k)}(t)}{k!}(z - t)^k,$$

для нее имеем $F(z) = 0$, $F(z_0) = r_n(z)$. Продифференцируем $F(t)$ по переменной t:

$$F'(t) = -\frac{f^{(n+1)}(t)}{n!}(z - t)^n.$$

В силу неравенства (9) и условия (11) получаем

$$\|r_n(z)\| = \|F(z_0) - F(z)\| \le 2 \max_{\zeta \in [z, z+h]} \|F'(\zeta)\| \cdot \|z_0 - z\| \le 2 \max_{\zeta \in [z, z+h]} \left\| \frac{f^{(n+1)}(t)}{n!}(z - t)^n \right\| \cdot \|z_0 - z\| \le$$

$$\le 2 \sup_{\zeta \in [z, z+h]} \frac{1}{n!} \|f^{(n+1)}(\zeta)\| \cdot \|(z - t)^n\| \cdot \|z_0 - z\| \le \frac{2}{n!} M e^{AR^m} R^{n+1} \xrightarrow[n \to \infty]{} 0$$

при условии $\|z - t\| \le R$ и $\|z - z_0\| \le R$. Отсюда выводим

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k,$$

где ряд сходится равномерно в круге $\|z - z_0\| \le R$. Теорема доказана.

Следствие. *Остаточный член формулы Тейлора в форме Пеано имеет вид*

$$r_n(z) = o\left(\|z - z_0\|^n\right), \; n \to \infty.$$

Определение 5. Функция f h-аналитична в области $D \in \mathbb{C}_h$, если она h-аналитична во всех точках этой области.

Пусть f h-аналитична в точке z_0, следовательно, в окрестности точки z_0 имеем

$$f(z) = c_k(z - z_0)^k + c_{k+1}(z - z_0)^{k+1} + \ldots, \tag{12}$$

где $c_k \ne 0$, $k \ge 0$.

Определение 6. Точка z_0 называется нулем порядка k функции f, если $k \ge 1$ в (12).

Из (12) вытекает представление

$$f(z) = (z - z_0)^k \varphi(z),$$

где $\varphi(z) = c_k + c_{k+1}(z - z_0) + \ldots$, $\varphi(z)$ h-аналитична в окрестности точки z_0, $\varphi(z_0) = c_k \ne 0$. В силу непрерывности функции $\varphi(z)$ существует окрестность $U(z_0): \varphi(z) \ne 0 \; \forall z \in U(z_0)$. Если c_k не является делителем нуля, то существует окрестность $V(z_0)$ такая, что в этой окрестности $f(z)$ не имеет других нулей, кроме точки z_0. Отсюда вытекает следующая теорема.

Теорема 8 (теорема единственности для h-аналитических функций). *Пусть f_1 и f_2 h-аналитичны в области $D \subset \mathbb{C}_h$, $f_1(z_k) \equiv f_2(z_k)$, где $z_k \in D$, $\lim\limits_{k \to \infty} z_k = z_0 \in D$, $(1 \pm j)(z_k - z_0) \ne 0$, $k = 1, 2, \ldots$. Тогда $f_1(z) \equiv f_2(z)$ всюду в D.*

Доказательство. Пусть $f(z) = f_1(z) - f_2(z)$. В некоторой окрестности $U(z_0)$

$$f(z) = c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + \ldots.$$

Имеем $f(z_k) = 0$, следовательно,

$$f(z_0) = \lim_{k \to \infty} f(z_k) = 0,$$

тогда $c_0 = 0$. Отсюда получаем

$$f(z) = (z - z_0)(c_1 + c_2(z - z_0) + \ldots) = (z - z_0)\varphi_1(z),$$

где $\varphi_1(z) = c_1 + c_2(z - z_0) + \ldots$.

Теперь

$$f(z_k) = (z_k - z_0)\varphi_1(z_k) = 0,$$

тогда $\varphi_1(z_k) = 0$ и $\varphi_1(z_0) = 0$, значит, $c_1 = 0$, следовательно,

$$\varphi_1(z) = c_2(z - z_0) + c_3(z - z_0)^2 + \ldots = (z - z_0)(c_2 + c_3(z - z_0) + \ldots) = (z - z_0)\varphi_2(z).$$

Поскольку

$$\varphi_1(z) = (z_k - z_0)\varphi_2(z) = 0,$$

то $\varphi_2(z_k) = 0$ и $\varphi_2(z_0) = 0$, а значит, $c_2 = 0$ и т. д., следовательно, $c_n = 0\ \forall n$. Таким образом, $f(z) \equiv 0$ $\forall z \in U(z_0)$. Пусть $M \subset D$ – множество нулей функции f, $\overset{\circ}{M}$ – его внутренность, причем $\overset{\circ}{M} \neq \varnothing$. Если $\overset{\circ}{M} = D$, то теорема доказана. В случае если $\overset{\circ}{M} \subsetneq D$, существует граничная точка d множества $\overset{\circ}{M}$, являющаяся внутренней точкой множества D. Также существует последовательность $d_n \in \overset{\circ}{M}$ такая, что

$$\lim_{n\to\infty} d_n = d,$$

причем $(1 \pm j)(d_n - d) \neq 0$,

$$f(d) = \lim_{n\to\infty} d_n = 0,$$

следовательно, существует окрестность $V(d)$ такая, что $f(z) \equiv 0\ \forall z \in V(d)$. Это противоречит тому, что d – граничная точка $\overset{\circ}{M}$. Теорема доказана.

Библиографические ссылки

1. Antonuccio F. Semi-complex analysis and mathematical physics [Internet]. 2008 [cited 2021 January 23]. 56 p. Available from: https://arxiv.org/abs/gr-qc/9311032.

2. Розенфельд БА. *Неевклидовы геометрии.* Москва: Государственное издательство технико-теоретической литературы; 1955. 744 с.

3. Ивлев ДД. О двойных числах и их функциях. В: Бронштейн ИН, Лопшиц АМ, Ляпунов АА, Маркушевич АИ, Яглом ИМ, редакторы. *Математика, ее преподавание, приложения и история*. Москва: Государственное издательство физико-математической литературы; 1961. с. 197–203. (Математическое просвещение; выпуск 6).

4. Deckelman S, Robson B. Split-complex numbers and Dirac brackets. *Communications in Information and Systems.* 2014;14(3): 135–159. DOI: 10.4310/CIS.2014.v14.n3.a1.

5. Khrennikov A. Hyperbolic quantum mechanics. *Advances in Applied Clifford Algebras.* 2003;13(1):1–9. DOI: 10.1007/s00006-003-0001-1.

6. Зверович ЭИ, Павловский ВА. Нахождение областей сходимости и вычисление сумм степенных рядов от *h*-комплексного переменного. *Весці Нацыянальнай акадэміі навук Беларусі. Серыя фізіка-матэматычных навук.* 2020;56(2):189–193. DOI: 10.29235/1561-2430-2020-56-2-189-193.

7. Павловский ВА. Алгебраические уравнения с вещественными коэффициентами в кольце *h*-комплексных чисел. *Весці БДПУ. Серыя 3. Фізіка. Матэматыка. Інфарматыка. Біялогія. Геаграфія.* 2020;4:25–31.

8. Зверович ЭИ. *Вещественный и комплексный анализ. Часть 3. Дифференциальное исчисление функций векторного аргумента.* Минск: Вышэйшая школа; 2006. 129 с.

References

1. Antonuccio F. Semi-complex analysis and mathematical physics [Internet]. 2008 [cited 2021 January 23]. 56 p. Available from: https://arxiv.org/abs/gr-qc/9311032.

2. Rosenfeld BA. *Neevklidovy geometrii* [Non-Euclidean geometries]. Moscow: Gosudarstvennoe izdatel'stvo tekhniko-teoreticheskoi literatury; 1955. 744 p. Russian.

3. Ivlev DD. [On double numbers and their functions]. In: Bronshtein IN, Lopshits AM, Lyapunov AA, Markushevich AI, Yaglom IM, editors. *Matematika, ee prepodavanie, prilozheniya i istoriya* [Mathematics, its teaching, applications and history]. Moscow: Gosudarstvennoe izdatel'stvo fiziko-matematicheskoi literatury; 1961. p. 197–203. (Matematicheskoe prosveshchenie; issue 6). Russian.

4. Deckelman S, Robson B. Split-complex numbers and Dirac brackets. *Communications in Information and Systems.* 2014;14(3): 135–159. DOI: 10.4310/CIS.2014.v14.n3.a1.

5. Khrennikov A. Hyperbolic quantum mechanics. *Advances in Applied Clifford Algebras.* 2003;13(1):1–9. DOI: 10.1007/s00006-003-0001-1.

6. Zverovich EI, Pavlovsky VA. Finding the areas of convergence and calculating sums of power series from an *h*-complex variable. *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics Series.* 2020;56(2):189–193. Russian. DOI: 10.29235/1561-2430-2020-56-2-189-193.

7. Pavlovsky VA. Algebraic equations with material coefficients in the ring of *h*-complex numbers. *Vesci BDPU. Seryja 3. Fizika. Matjematyka. Infarmatyka. Bijalogija. Geagrafija.* 2020;4:25–31. Russian.

8. Zverovich EI. *Veshchestvennyi i kompleksnyi analiz. Chast' 3. Differentsial'noe ischislenie funktsii vektornogo argumenta* [Real and complex analysis. Part 3. Differential calculus of vector argument functions]. Minsk: Vyshjejshaja shkola; 2006. 129 p. Russian.

Получена 23.03.2021 / исправлена 04.06.2021 / принята 04.06.2021.
Received 23.03.2021 / revised 04.06.2021 / accepted 04.06.2021.

Дифференциальные уравнения и оптимальное управление

Differential equations and optimal control

УДК 517.977.5

МЕТОД ПОСТРОЕНИЯ ОПТИМАЛЬНОЙ СТРАТЕГИИ УПРАВЛЕНИЯ В ЛИНЕЙНОЙ ТЕРМИНАЛЬНОЙ ЗАДАЧЕ

Д. А. КОСТЮКЕВИЧ[1)]***, Н. М. ДМИТРУК***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается задача оптимального управления линейной дискретной системой с неизвестными ограниченными возмущениями, которую требуется за конечное время перевести с гарантией на терминальное множество, обеспечивая при этом минимум гарантированного значения терминального критерия качества. Определяется оптимальная стратегия управления, учитывающая информацию о состоянии системы в один будущий момент времени, и предлагается эффективный метод ее вычисления. Результаты численных экспериментов демонстрируют улучшение качества управления на основе введенной оптимальной стратегии в сопоставлении с оптимальной гарантирующей программой при сравнимой трудоемкости их вычисления.

Ключевые слова: линейная система; возмущения; оптимальное управление; стратегия управления; алгоритм.

Образец цитирования:
Костюкевич ДА, Дмитрук НМ. Метод построения оптимальной стратегии управления в линейной терминальной задаче. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:38–50 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-38-50

For citation:
Kastsiukevich DA, Dmitruk NM. A method for constructing an optimal control strategy in a linear terminal problem. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:38–50.
https://doi.org/10.33581/2520-6508-2021-2-38-50

Авторы:
Дмитрий Аркадьевич Костюкевич – старший преподаватель кафедры методов оптимального управления факультета прикладной математики и информатики.
Наталия Михайловна Дмитрук – кандидат физико-математических наук, доцент; заведующий кафедрой методов оптимального управления факультета прикладной математики и информатики.

Authors:
Dzmitry A. Kastsiukevich, senior lecturer at the department of optimal control methods, faculty of applied mathematics and computer science.
kostukda@bsu.by
https://orcid.org/0000-0002-5803-9800
Natalia M. Dmitruk, PhD (physics and mathematics), docent; head of the department of optimal control methods, faculty of applied mathematics and computer science.
dmitrukn@bsu.by
https://orcid.org/0000-0003-1845-4927

A METHOD FOR CONSTRUCTING AN OPTIMAL CONTROL STRATEGY IN A LINEAR TERMINAL PROBLEM

D. A. KASTSIUKEVICH [a], ***N. M. DMITRUK*** [a]

[a] *Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: D. A. Kastsiukevich (kostukda@bsu.by)

This paper deals with an optimal control problem for a linear discrete system subject to unknown bounded disturbances, where the control goal is to steer the system with guarantees into a given terminal set while minimising the terminal cost function. We define an optimal control strategy which takes into account the state of the system at one future time instant and propose an efficient numerical method for its construction. The results of numerical experiments show an improvement in performance under the optimal control strategy in comparison to the optimal open-loop worst-case control while maintaining comparable computation times.

Keywords: linear system; disturbance; optimal control; control strategy; algorithm.

Introduction

Optimal control problems for dynamical systems under uncertainty have been studied in the literature since late 1960s [1–3]. The simplest approach that guarantees constraints satisfaction and achieves the guaranteed value of the cost at the worst-case disturbance realisation is to find an optimal open-loop worst-case control. The optimal open-loop worst-case control is constructed before the control process starts and is not corrected during it; no information about possible future state measurements is used for its construction. It is well known that optimal open-loop worst-case controls underestimate the potential of the control process, i. e., they give a conservative estimate of the guaranteed optimal value of the problem and often cannot be constructed because of the constraints infeasibility (see, e. g., [4–6]). However, dynamic programming takes into account all future state realisations, but the practical derivation of the dynamic programming strategy is computationally intense with the exception of special cases of low dimensional systems and short control intervals.

Therefore, such control strategies are relevant that take into account some information about the future states of the system and at the same time the complexity of their construction is comparable to the complexity of calculating optimal open-loop worst-case controls. One of the possible approaches was proposed in papers [6–8]. In [6] linear terminal problems were considered [7] deals with linear-quadratic optimal control problems and [8] deals with problems of minimising the total momentum of the control input. All these papers assume that before control process starts, we can choose one or more time instants (closing time instants of the system according to [6; 8]), at which we can measure exactly the system state and make corrections in the control input.

This paper deals with the problem considered in [6]. In contrast to [6], where a complex iterative algorithm was used to construct an optimal control strategy with one closing instant, which requires sequential optimisation first in control inputs and then in a parameter, we use the ideas of [8] to reduce the problem under consideration to a single linear program, which allows to calculate the optimal control and the optimal parameter simultaneously.

Compared to [8], the problem studied in this paper has a terminal performance index and a discrete time system, while in [8] a Lagrange cost of a special type and continuous time systems are investigated. Further comparison of the results from [8] and the ones of this paper, the drawbacks and advantages of the two methods are discussed in example 2. Two more examples demonstrate the efficiency of the new approach.

Optimal open-loop worst-case control

Consider a linear discrete-time time-invariant control system with a disturbance

$$x(t+1) = Ax(t) + Bu(t) + Mw(t),\ x(0) = x_0,\ t = 0, 1, \ldots, T-1, \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in U \subset \mathbb{R}^r$ is the control input, $w(t) \in W \subset \mathbb{R}^p$ is the unknown disturbance at time t, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times r}$, $M \in \mathbb{R}^{n\times p}$ are given matrices; $U = \left\{u \in \mathbb{R}^r : u_{\min} \le u \le u_{\max}\right\}$, $W = \left\{w \in \mathbb{R}^p : \|w\|_\infty \le \right.$ $\left. \le w_{\max}\right\}$, where $u_{\min}, u_{\max} \in \mathbb{R}^r$, $w_{\max} > 0$, $\|z\|_\infty = \max\limits_i |z_i|$. A trajectory of system (1) generated by a feasible control input $u(\cdot) = \left(u(t) \in U,\ t = 0, 1, \ldots, T-1\right)$ and a disturbance $w(\cdot) = \left(w(t) \in W,\ t = 0, 1, \ldots, T-1\right)$ is denoted by $x\left(t \middle| x_0, u(\cdot), w(\cdot)\right)$, $t = 0, 1, \ldots, T-1$.

Given a terminal set $X_T = \left\{x \in \mathbb{R}^n : g_{\min} \le Hx \le g_{\max}\right\}$, where $H \in \mathbb{R}^{m \times n}$, $g_{\min}, g_{\max} \in \mathbb{R}^m$, the control goal is to steer system (1) at time instant T to the terminal set with guarantees. The requirement of a guaranteed (robust) steering to the set X_T without any further assumptions yields the definition of a feasible open-loop worst-case control.

Definition 1. A control input $u(\cdot)$ is called a feasible open-loop worst-case control if for any possible realisation of the disturbance $w(\cdot)$ it steers system (1) at time instant T into the terminal set, i. e. the following inclusion holds

$$x\left(T \,|\, x_0,\, u(\cdot),\, w(\cdot)\right) \in X_T \;\; \forall\, w(t) \in W,\; t = 0,\, 1,\, \ldots,\, T-1. \tag{2}$$

The quality of a feasible open-loop worst-case control $u(\cdot)$ is measured by the value

$$J(u) = \max_{w(\cdot)} c' x(T),\; c \in \mathbb{R}^n, \tag{3}$$

that represents the terminal cost (Mayer's performance index) at the worst realisation of the disturbance and is called the guaranteed value of the terminal cost. Here the prime symbol denotes a transpose.

Definition 2. A feasible open-loop worst-case control $u^0(\cdot)$ is called optimal if it minimises the guaranteed value of the terminal cost (3): $J\left(u^0\right) = \min_u J(u)$.

In guaranteed (robust) optimal control problems, along with the disturbed system (1), one considers a so-called nominal system

$$x_0(t+1) = Ax_0(t) + Bu(t),\; x_0(0) = x_0,\; t = 0,\, 1,\, \ldots,\, T-1,$$

that is used to formulate a deterministic optimal control problem equivalent to the problem of minimising the cost (3) subject to system (1) and inclusion (2). The method for constructing this deterministic problem is well investigated in the literature (see, e. g., [6]). It uses the linearity of system (1) and estimates of the worst-case realisations of disturbances in the directions specified by the vector c and the rows h_i' of the matrix H:

$$\gamma_0(\tau) = w_{\max} \sum_{t=0}^{T-\tau-1} \left\| c'A^t M \right\|_1,\; \gamma_i(\tau) = w_{\max} \sum_{t=0}^{T-\tau-1} \left\| h_i'A^t M \right\|_1,\; i = 1,\, \ldots,\, m,\; \|z\|_1 = \sum_i |z_i|.$$

The vector of estimates $\gamma(0) = \left(\gamma_i(0),\; i = 1,\, \ldots,\, m\right)$ allows to define a «tightened» terminal set for the nominal system and to formulate the deterministic problem in the form

$$J(u) = \min_{u(\cdot)} c' x_0(T) + \gamma_0(0), \tag{4}$$

$$x_0(t+1) = Ax_0(t) + Bu(t),\;\; x_0(0) = x_0,\;\; u(t) \in U,\; t = 0,\, 1,\, \ldots,\, T-1,$$

$$g_{\min} + \gamma(0) \le Hx_0(T) \le g_{\max} - \gamma(0).$$

Using the formula $x_0(T) = A^T x_0 + \sum_{t=0}^{T-1} A^{T-t-1} Bu(t)$ for the terminal state of the nominal system and substituting it in problem (4) we conclude that the optimal open-loop worst-case control $u^0(\cdot)$ can be calculated as a solution to the linear program

$$\min_{u(\cdot)} \sum_{t=0}^{T-1} c' A^{T-t-1} Bu(t),$$

$$g_{\min} + \gamma(0) - HA^T x_0 \le \sum_{t=0}^{T-1} HA^{T-t-1} Bu(t) \le g_{\max} - \gamma(0) - HA^T x_0,$$

$$u_{\min} \le u(t) \le u_{\max},\; t = 0,\, \ldots,\, T-1,$$

where the constant $c'A^T x_0 + \gamma_0(0)$ in the cost is omitted.

The optimal open-loop worst-case control is the simplest solution of the problem under consideration, when system (1) has to be robustly steered to the terminal set while minimising the terminal cost (3). The open-loop control does not take into account the possibility of future state measurements of system (1), that allow to close the control loop and to make corrections to the planned control inputs (see, e. g., [7; 8]). In contrast to optimal open-loop worst-case controls, such a possibility is taken into account by the control strategies. One of such control strategies is introduced in the next section.

Optimal control strategy

Before the control process starts, we fix a time instant $T_1 \in \{1, 2, \ldots, T-1\}$ that is referred to as the closing instant of system (1) (see [6; 8]). Denote $\Delta_0 = \{0, 1, \ldots, T_1 - 1\}$, $\Delta_1 = \{T_1, T_1 + 1, \ldots, T-1\}$; $u_k(\cdot) = \big(u_k(t) \in U, t \in \Delta_k\big)$ is the control input on the interval Δ_k; $w_k(\cdot) = \big(w_k(t) \in W, t \in \Delta_k\big)$ is the disturbance on Δ_k; $U_k = \{u_k(\cdot) : u_k(t) \in U, t \in \Delta_k\}$ is the set of feasible control inputs on Δ_k; $W_k = \{w_k(\cdot) : w_k(t) \in W, t \in \Delta_k\}$ is the set of possible disturbances on Δ_k, $k = 0, 1$.

Assume that on the interval Δ_0 a control input $u_0(\cdot) = u_0\big(\cdot \big| x_0\big) \in U_0$ is chosen. At time T_1 system (1) reaches a state x_1 that belongs to the set

$$X\big(T_1 \big| x_0, u_0(\cdot)\big) = \big\{x \in \mathbb{R}^n : x = x\big(T_1 \big| x_0, u_0(\cdot), w_0(\cdot)\big), w_0(\cdot) \in W_0\big\}.$$

Following [7; 8], it is assumed that at time instant T_1 we can:

1) measure exactly the current state $x_1 = x\big(T_1 \big| x_0, u_0(\cdot), w_0(\cdot)\big)$;

2) choose a new control input $u_1(\cdot) = u_1\big(\cdot \big| x_1\big) \in U_1$ on Δ_1 taking into account the obtained state measurement x_1.

Taking into account 1) and 2) we look for a solution of the problem under consideration in terms of a control strategy (with the closing instant T_1):

$$\pi_1 = \big\{u_0\big(\cdot \big| x_0\big); u_1\big(\cdot \big| x_1\big), x_1 \in X\big(T_1 \big| x_0, u_0\big(\cdot \big| x_0\big)\big)\big\},$$

where the control input $u_0(\cdot) = u_0\big(\cdot \big| x_0\big)$ is referred to as an initial open-loop control.

A trajectory of control system (1), corresponding to a strategy π_1 and a disturbance $w(\cdot) = \big(w_0(\cdot), w_1(\cdot)\big)$, is defined as a sequential solution of two systems [7; 8]:

$$x(t+1) = Ax(t) + Bu_0(t) + Mw_0(t), \; x(0) = x_0, \; t \in \Delta_0,$$

$$x(t+1) = Ax(t) + Bu_1\big(t \big| x\big(T_1\big)\big) + Mw_1(t), \; x\big(T_1\big) = x\big(T_1 \big| x_0, u_0(\cdot), w_0(\cdot)\big), \; t \in \Delta_1.$$

Now we discuss conditions for the strategy π_1 to be feasible, i. e. for the trajectory defined above to guarantee the terminal constraints satisfaction.

First, the control input $u_1\big(\cdot \big| x_1\big) \in U_1$ that is chosen at the time instant T_1, must satisfy the inclusion

$$X\big(T \big| x_1, u_1\big(\cdot \big| x_1\big)\big) \subseteq X_T, \tag{5}$$

where $X\big(T \big| x_1, u_1(\cdot)\big) = \big\{x \in \mathbb{R}^n : x = x\big(T \big| x_1, u_1(\cdot), w_1(\cdot)\big), w_1(\cdot) \in W_1\big\}$ is the set of possible terminal states $x\big(T \big| x_1, u_1(\cdot), w_1(\cdot)\big)$ of system (1) with the initial condition $x\big(T_1\big) = x_1$, the control input $u_1(\cdot)$ and the disturbance $w_1(\cdot)$.

Secondly, the control input $u_0(\cdot)$ should be such that for all states x_1 from the set $X\big(T_1 \big| x_0, u_0(\cdot)\big)$ there exist a control $u_1\big(\cdot \big| x_1\big)$, satisfying (5). Summarising, we obtain the next definition.

Definition 3. A strategy π_1 is called a feasible control strategy if

$$X\big(T \big| x_1, u_1\big(\cdot \big| x_1\big)\big) \subseteq X_T \quad \forall x_1 \in X\big(T_1 \big| x_0, u_0(\cdot)\big).$$

Obviously, an arbitrary feasible strategy with the initial open-loop control $u_0\big(\cdot \big| x_0\big)$ is not better than a feasible control strategy of the form

$$\pi_1 = \big\{u_0\big(\cdot \big| x_0\big); u_1^0\big(\cdot \big| x_1\big), x_1 \in X\big(T_1 \big| x_0, u_0\big(\cdot \big| x_0\big)\big)\big\}, \tag{6}$$

which on Δ_1 consists of the optimal open-loop worst-case controls $u_1^0\big(\cdot \big| x_1\big)$ for states x_1. Every open-loop control $u_1^0\big(\cdot \big| x_1\big)$ for a fixed x_1 is the solution of the problem

$$J_1\big(x_1\big) = \min_{u_1(\cdot) \in U_1} \max_{w_1(\cdot) \in W_1} c'x\big(T \big| x_1, u_1(\cdot), w_1(\cdot)\big) \tag{7}$$

subject to (5). The quality of strategy (6) is obviously measured by the value

$$V\big(\pi_1\big) = \max_{w_0(\cdot) \in W_0} J_1\big(x\big(T_1 \big| x_0, u_0(\cdot), w_0(\cdot)\big)\big).$$

Note that if problem (7) is infeasible, then we assume $J_1(x_1) = +\infty$. Therefore, if the strategy π_1 is not feasible, i. e. for some $x_1 \in X(T_1|x_0, u_0(\cdot))$ there is no control input $u_1(\cdot|x_1)$ that satisfies (5), then $V(\pi_1) = +\infty$.

Definition 4. A feasible control strategy

$$\pi_1^0 = \left\{u_0^0(\cdot|x_0);\, u_1^0(\cdot|x_1),\, x_1 \in X\left(T_1|x_0,\, u_0^0(\cdot|x_0)\right)\right\}, \tag{8}$$

is called optimal, if $V(\pi_1^0) = \min V(\pi_1)$, where minimum is taken over all feasible strategies of the form (6). A control input $u_0^0(\cdot|x_0)$ is called the optimal initial open-loop control (within the optimal strategy π_1^0).

Hence, the control strategy (8) is optimal if $u_0^0(\cdot|x_0)$ is a solution of the minimax problem

$$\begin{gathered} V(\pi_1^0) = \min_{u_0(\cdot)\in U_0} \max_{w_0(\cdot)\in W_0} J_1(x(T_1)), \\ x(t+1) = Ax(t) + Bu_0(t) + Mw_0(t),\ x(0) = x_0,\ t \in \Delta_0, \end{gathered} \tag{9}$$

and $u_1^0(\cdot|x_1)$ are solutions of problems (7) for the states $x_1 \in X\left(T_1|x_0,\, u_0^0(\cdot|x_0)\right)$.

Problem (9) implies that the optimal guaranteed value of the strategy π_1^0 is equal to

$$V(\pi_1^0) = \min_{u_0(\cdot)\in U_0} \max_{w_0(\cdot)\in W_0} \min_{u_1(\cdot)\in U_1} \max_{w_1(\cdot)\in W_1} c'x\left(T|x\left(T_1|x_0,\, u_0(\cdot),\, w_0(\cdot)\right),\, u_1\left(\cdot|x\left(T_1|x_0,\, u_0(\cdot),\, w_0(\cdot)\right)\right),\, w_1(\cdot)\right),$$

while the optimal guaranteed value of the open-loop worst-case control $u^0(t)$ is calculated as

$$J(u^0) = \min_{u_0(\cdot)\in U_0} \min_{u_1(\cdot)\in U_1} \max_{w_0(\cdot)\in W_0} \max_{w_1(\cdot)\in W_1} c'x\left(T|x\left(T_1|x_0,\, u_0(\cdot),\, w_0(\cdot)\right),\, u_1(\cdot),\, w_1(\cdot)\right).$$

Taking into account the minimax inequality we conclude that $V(\pi_1^0) \le J(u^0)$. In the last section we will provide some examples where the optimal control strategy with one closing instant achieves a significant improvement in comparison to the optimal open-loop worst-case control.

Calculating the optimal initial open-loop control

Before the control process starts, we need to know only the optimal initial open-loop control $u_0^0(\cdot|x_0)$. The collection of the optimal open-loop worst-case controls $u_1^0(\cdot|x_1)$ is not calculated in advance. The control input $u_1^0(\cdot|x_1(T_1))$ is only found at the closing time instant T_1, when the current state $x(T_1)$ is measured. Therefore, the purpose of this section is to propose an efficient method for calculating the optimal initial open-loop control $u_0^0(\cdot|x_0)$, i. e. solving problem (9).

Problem (9) is the terminal control problem of the same type as the problem for calculating the optimal open-loop worst-case control $u^0(\cdot)$. It has no explicit terminal constraints; however, these are implicitly imposed by the condition $x(T_1) \in X_1 = \{x_1 : J_1(x_1) < +\infty\}$. The principal difficulty in solving problem (9) is that the function J_1 in the performance index is defined implicitly as the optimal value of problem (7). In this regard, for the purposes of further presentation, we reformulate problem (9) in an equivalent form (see [9]):

$$\begin{gathered} V(\pi_1^0) = \min_{u_0(\cdot),\,\alpha} \alpha, \\ x(t+1) = Ax(t) + Bu_0(t) + Mw_0(t),\ x(0) = x_0,\ u_0(t) \in U_0,\ t \in \Delta_0, \\ J_1(x(T_1)) \le \alpha\ \forall\, w_0(\cdot) \in W_0. \end{gathered} \tag{10}$$

The function $J_1(x_1)$, $x_1 \in \mathbb{R}^n$, as the optimal value of a linear program (to which problem (7) is reduced), is a piecewise linear convex function (see [10, p. 180]), therefore for any fixed $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ the α-level set $X_1(\alpha) = \{x_1 \in X_1 : J_1(x_1) \le \alpha\}$ is a convex polyhedron. Here $\alpha_{\min} = \inf c'x + \gamma_0(T_1)$, $\alpha_{\max} = \sup c'x - \gamma_0(T_1)$, subject to $g_{\min} + \gamma(T_1) \le Hx \le g_{\max} - \gamma(T_1)$. Then in (10) the terminal constraint has the form $x(T_1) \in X_1(\alpha)$ $\forall\, w_0(\cdot) \in W_0$.

Since the exact description of the polyhedra $X_1(\alpha)$ for all values of the parameter α is difficult, in [6; 8] it was proposed to replace $X_1(\alpha)$ with their outer polyhedral approximations with normals to the faces of these

polyhedra being independent of α. Let $p_j \in \mathbb{R}^n$, $j = 1, 2, \ldots, m_1$, $\|p_j\| = 1$, be a collection of vectors which represent the mentioned normals, and $P_1 \in \mathbb{R}^{m_1 \times n}$ be a matrix, which rows are the vectors p_j. Denote

$$f_j(\alpha) = \max p_j' x_1,\ x_1 \in X_1(\alpha). \tag{11}$$

Then the outer approximating polyhedron for $X_1(\alpha)$ is $\bar{X}_1(\alpha) = \{x_1 \in \mathbb{R}^n : P_1 x_1 \le f(\alpha)\}$, where $f(\alpha) = (f_j(\alpha),\ j = 1, \ldots, m_1)$. With a sufficiently large set of vectors p_j, $j = 1, 2, \ldots, m_1$, $\bar{X}_1(\alpha)$ approximate the sets $X_1(\alpha)$, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, quite accurately.

In [6], in order to solve problem (10), an iterative algorithm was proposed. Each iteration of the algorithm for the current value α_k refines the approximation $\bar{X}_1(\alpha_k)$ and calculates the control $u^0_{\alpha_k}(t)$, $t \in \Delta_0$, that guarantees steering the system to the set $\bar{X}_1(\alpha_k)$ at the time instant T_1. Thus, at each iteration k the algorithm solves $m_1(k)$ problems (11) and one control problem. The value α_{k+1} is found by any line search method.

In what follows we propose a method for solving problem (10), which does not require application of the iterative procedure described in [6].

Denote $G_1 = HA^{T-T_1}$, $c_1' = c'A^{T-T_1}$ and consider the case when $\operatorname{rank}(G_1' | c_1) = m + 1 \le n$. Simple arguments yield that in this case $\alpha_{\min} = -\infty$, $\alpha_{\max} = +\infty$, i. e. function (11) is defined on the entire real axis. Let us show that $f(\alpha)$, $\alpha \in \mathbb{R}$, is affine.

Assumption 1. For any $j = 1, 2, \ldots, m_1$ the vector p_j is such that equalities $\operatorname{rank}(G_1' | c_1) = \operatorname{rank}(G_1' | c_1 | p_j) = m + 1$ hold.

Proposition. *Let assumption 1 hold. Then*

$$f(\alpha) = f_0 + \lambda\alpha, \tag{12}$$

where $f_0 = f(0)$, $\lambda = (\lambda_j,\ j = 1, \ldots, m_1)$, λ_j *satisfies the conditions* $G_1' y + c_1 \lambda_j = p_j$, $\lambda_j \ge 0$.

P r o o f. Consider problem (11) for a fixed index j. The set $X_1(\alpha)$ consists of those and only those vectors x_1 for which the following system is feasible

$$x(t+1) = Ax(t) + Bu_1(t),\ x(T_1) = x_1,\ u_1(t) \in U,\ t \in \Delta_1,$$

$$g_{\min} + \gamma(T_1) \le Hx(T) \le g_{\max} - \gamma(T_1),\ c'x(T) \le \alpha - \gamma_0(T_1).$$

Following the arguments that were used to reduce the problem for constructing the optimal open-loop worst-case control to problem (4), problem (11) can also be reduced to a linear program. We represent it in the form

$$\begin{gathered} f_j(\alpha) = \max_{x_1,\, u_1} p_j' x_1, \\ g_{\min} + \gamma(T_1) \le G_1 x_1 + \sum_{t \in \Delta_1} G_1 D(t) u_1(t) \le g_{\max} - \gamma(T_1), \\ c_1' x_1 + \sum_{t \in \Delta_1} c_1' D(t) u_1(t) \le \alpha - \gamma_0(T_1), \\ u_{\min} \le u_1(t) \le u_{\max},\ t \in \Delta_1, \end{gathered} \tag{13}$$

where $D(t) = A^{T_1 - t - 1} B$, $t \in \Delta_1$.

The problem dual to (13) has the form

$$\begin{gathered} f_j(\alpha) = \min_{\substack{y^*,\, y_*,\, \lambda_j, \\ v_*(t),\, v^*(t),\, t \in \Delta_1}} \left(g_{\max} - \gamma(T_1)\right)' y^* - \left(g_{\min} + \gamma(T_1)\right)' y_* + \left(\alpha - \gamma_0(T_1)\right)\lambda_j + \sum_{t \in \Delta_1} \left(u_{\max}' v^*(t) - u_{\min}' v_*(t)\right), \\ G_1' y^* - G_1' y_* + c_1 \lambda_j = p_j, \\ D(t)' G_1' y^* - D(t)' G_1' y_* + D(t)' c_1 \lambda_j + v^*(t) - v_*(t) = 0, \\ \lambda_j \ge 0,\ y^* \ge 0,\ y_* \ge 0,\ v^*(t) \ge 0,\ v_*(t) \ge 0,\ t \in \Delta_1, \end{gathered} \tag{14}$$

and the complimentary slackness conditions hold [9; 10]. Note that all the dual variables in (14), similarly to λ_j, depend on the index j that is omitted for simplicity of presentation.

Denote $y = y^* - y_*$. Then, according to assumption 1, the system of linear algebraic equations (14) has a unique solution (y, λ_j). If that solution has $\lambda_j < 0$, then the dual problem (14) is infeasible and the primal problem (13) is unbounded on $X_1(\alpha)$. Let $\lambda_j \geq 0$, then both problems (13), (14) are feasible. The second group of equality constraints in problem (14) can be represented as

$$v^*(t) - v_*(t) = -D(t)' p_j,\ t \in \Delta_1.$$

Taking into account non-negativeness of the dual variables and the complementary slackness conditions, it is clear that the dual variables are calculated according to the formulae

$$y_i^* = \begin{cases} y_i, & y_i \geq 0, \\ 0, & y_i < 0, \end{cases} \quad y_{*i} = \begin{cases} 0, & y_i \geq 0, \\ -y_i, & y_i < 0, \end{cases} \quad i = 1, \ldots, m,$$

$$v_*(t) = \left(v_{*k}(t),\ k = 1, \ldots, r\right),\ v^*(t) = \left(v_k^*(t),\ k = 1, \ldots, r\right), \tag{15}$$

$$v_{*k}(t) = \begin{cases} v_k(t), & v_k(t) \geq 0, \\ 0, & v_k(t) < 0, \end{cases} \quad v_k^*(t) = \begin{cases} 0, & v_k(t) \geq 0, \\ -v_k(t), & v_k(t) < 0, \end{cases} \quad t \in \Delta_1,$$

where $v(t) = \left(v_k(t),\ k = 1, \ldots, r\right) = D(t)' p_j,\ t \in \Delta_1$.

Problem (14) for $\alpha = 0$ has same optimal solution, therefore we can conclude that $f_j(\alpha) = f_j(0) + \lambda_j \alpha$, which proves formula (12).

Along with the proposition we derived a simple formula for calculating the components of the vector f_0 that allows to avoid solving linear programs (13) or (14):

$$f_{0j} = f_j(0) = \left(g_{\max} - \gamma(T_1)\right)' y^* - \left(g_{\min} + \gamma(T_1)\right)' y_* - \gamma_0(T_1)\lambda_j + \sum_{t \in \Delta_1} \left(u'_{\max} v^*(t) - u'_{\min} v_*(t)\right).$$

It also follows that if $\operatorname{rank}\left(G_1' | c_1\right) \neq \operatorname{rank}\left(G_1' | c_1 | p_j\right)$, then $f_j(\alpha) = +\infty$.

Considering the relation (12) the problem for constructing the optimal initial open-loop control (10) (case $\operatorname{rank}\left(G_1' | c_1\right) = m + 1 \leq n$) can be presented in the form

$$V\left(\pi_1^0\right) = \min_{u_0(\cdot),\, \alpha} \alpha,$$

$$x(t+1) = Ax(t) + Bu_0(t) + Mw_0(t),\ x(0) = x_0,\ u_0(t) \in U,\ t \in \Delta_0, \tag{16}$$

$$P_1 x(T_1) \leq f_0 + \lambda\alpha\ \forall w_0(\cdot) \in W_0,$$

and further reduces to the linear program

$$\min_{u_0(\cdot),\, \alpha} \alpha,$$

$$\sum_{t \in \Delta_0} P_1 A^{T_1 - t - 1} B u_0(t) - \lambda\alpha \leq f_0 - \overline{\gamma} - P_1 A^{T_1} x_0, \tag{17}$$

$$u_{\min} \leq u_0(t) \leq u_{\max},\ t \in \Delta_0,$$

where $\overline{\gamma} = \left(\overline{\gamma}_j,\ j = 1, \ldots, m_1\right)$: $\overline{\gamma}_j = w_{\max} \sum_{t \in \Delta_0} \left\| p_j' A^t M \right\|_1$.

Now let us consider the case when $\operatorname{rank}\left(G_1' | c_1\right) = n < m + 1$. In this case $-\infty < \alpha_{\min} \leq \alpha_{\max} < +\infty$, function $f(\alpha)$, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, is piecewise affine and concave. Here in order to construct the optimal initial open-loop control $u_0^0(\cdot | x_0)$ we propose to replace the approximation of the set $X_1(\alpha)$ with the approximation of the set

$$\Xi_1(\alpha) = \left\{(\xi_0, \xi) \in \mathbb{R}^{m+1} : \exists u_1(\cdot) \in U_1,\ g_{\min} + \gamma(T_1) \leq \xi + \sum_{t \in \Delta_1} G_1 D(t) u_1(t) \leq g_{\max} - \gamma(T_1),\right.$$

$$\xi_0 + \sum_{t \in \Delta_1} c_1' D(t) u_1(t) \le \alpha - \gamma_0(T_1)\Big\},$$

by the polytope $\Xi_1(\alpha) = \Big\{(\xi_0, \xi) \in \mathbb{R}^{m+1} : q_0\xi_0 + Q_1\xi \le \overline{f}(\alpha)\Big\}$, where $q_0 = \big(q_{0j} \ge 0,\ j = 1, \ldots, m_1\big)$, matrix $Q_1 \in \mathbb{R}^{m_1 \times m}$ consists of rows $q_j' \in \mathbb{R}^m$, $j = 1, \ldots, m_1$, that are chosen in advance, and $\overline{f}(\alpha) = \big(\overline{f}_j(\alpha),\ j = 1, \ldots, m_1\big)$, where $\overline{f}_j(\alpha)$ is the optimal value of the linear program

$$\overline{f}_j(\alpha) = \max_{\xi_0,\, \xi,\, u_1} q_{0j}\xi_0 + q_j'\xi,$$

$$g_{\min} + \gamma(T_1) \le \xi + \sum_{t \in \Delta_1} G_1 D(t) u_1(t) \le g_{\max} - \gamma(T_1),\ \xi_0 + \sum_{t \in \Delta_1} c_1' D(t) u_1(t) \le \alpha - \gamma_0(T_1), \tag{18}$$

$$u_{\min} \le u_1(t) \le u_{\max},\ t \in \Delta_1.$$

Following the arguments of the proposition, we derive that $\overline{f}(\alpha),\ \alpha \in \mathbb{R}$, is calculated by the formulae

$$\overline{f}(\alpha) = \overline{f}_0 + \lambda\alpha,\ \lambda = q_0 \ge 0,$$

$$\overline{f}_j(0) = \big(g_{\max} - \gamma(T_1)\big)' y^* - \big(g_{\min} + \gamma(T_1)\big)' y_* - \gamma_0(T_1) q_{0j} + \sum_{t \in \Delta_1} \big(u_{\max}' v^*(t) - u_{\min}' v_*(t)\big),$$

where $y_*,\ y^*,\ v_*(t),\ v^*(t),\ t \in \Delta_1$, are found from (15) with the following adjustment: $y = q_j,\ v(t) = D'(t)G_1'q_j + D'(t)c_1 q_{0j},\ t \in \Delta_1$.

The problem for constructing the optimal initial open-loop control (10) (case $\operatorname{rank}\big(G_1' | c_1\big) = n < m + 1$) has the form

$$V\big(\pi_1^0\big) = \min_{u_0(\cdot),\, \alpha} \alpha,$$

$$x(t+1) = Ax(t) + Bu_0(t) + Mw_0(t),\ x(0) = x_0,\ u_0(t) \in U,\ t \in \Delta_0, \tag{19}$$

$$\overline{P}_1 x(T_1) \le \overline{f}_0 + q_0\alpha\ \forall w_0(\cdot) \in W_0$$

with $\overline{P}_1 = q_0 c_1' + Q_1 G_1$ and can be reduced to the linear program similarly to the reduction of problem (16) to problem (17).

The resulting linear program of the form (17) has $T_1 + 1$ variables and m_1 constraints. Depending on the required accuracy of approximation of the set $X_1(\alpha)$ or the set $\Xi_1(\alpha)$ the number of constraints can be quite large. However, in contrast to the method from [6], where first problems (13), (18) are solved and then the problem of the dimension comparable with the dimension of problems (17), (19) for a fixed parameter α is solved, problem (17) is solved only once and its solution immediately yields the optimal value of the parameter $\alpha^0 = V\big(\pi_1^0\big)$ and the optimal initial open-loop control $u_0^0(\cdot | x_0)$.

Note that in the second case $(\operatorname{rank}\big(G_1' | c_1\big) = n < m + 1)$ the space dimension where we approximate the set $\Xi_1(\alpha)$ is higher than the state space dimension n, which can be undesirable and lead to a significant increase in the number of constraints in problem (19). To avoid this problem one has to explore the piecewise affine structure of the function $f(\alpha),\ \alpha \in [\alpha_{\min}, \alpha_{\max}]$. This will be the focus of a future work.

It is also worth mentioning that the idea of approximating the set $\Xi_1(\alpha)$ can be applied in the case $\operatorname{rank}\big(G_1' | c_1\big) = m + 1$, if the number of terminal constraints m is less than the number of states n of the control system. Such an approach reduces the dimension of the space where approximations are constructed and is applied to solve examples 2 and 3 in the next section.

Examples

Let us illustrate the proposed method for constructing the optimal control strategy by three examples. The first example is a discrete analogue of the problem from [6], the second is the problem of minimising the total momentum of the control input from [8], and the third is a modification of the latter. Discrete systems for the examples are obtained by discretisation of continuous systems with the sampling period $h = 0,1$.

Example 1. Consider a discretised problem from [6]:

$$x_2(T) \to \max,$$

$$x(t+1) = \begin{pmatrix} 0.9950 & 0.0998 \\ -0.0998 & 0.9950 \end{pmatrix} x(t) + \begin{pmatrix} 0.0050 \\ 0.0998 \end{pmatrix} u(t) + \begin{pmatrix} 0.0050 \\ 0.0998 \end{pmatrix} w(t),\ x(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{20}$$

$$x(T) \in X_T = \left\{ x \in \mathbb{R}^2 : x_* \le x_1 \le x^* \right\},\ |u(t)| \le 1,\ |w(t)| \le 0.5,\ t = 0, \ldots, T-1.$$

Let us choose the control horizon $T = 120$ and the closing instant $T_1 = 80$. In [6] $x_* = 2$, $x^* = 7$; however, in this case there exists no feasible open-loop worst-case control (in both continuous problem from [6] and discrete problem (20)). Therefore, we choose $x_* = 2$, $x^* = 10$. This modification does not affect the optimal control strategy since the constraint $x_1(T) \le 10$ is not active, but it allows to compare the optimal open-loop worst-case control and the optimal strategy. In particular, problem (20) has the optimal open-loop worst-case control $u^0(\cdot)$ that achieves the optimal guaranteed value equal to $J(u^0) = 1.501102$, while the optimal control strategy (8) has the optimal guaranteed value $V(\pi_1^0) = \alpha^0 = 2.754\,215$. Calculating the optimal open-loop worst-case control takes 0.0151 s, while to obtain the optimal strategy we needed 0.0186 s. Figures 1 and 2 illustrate the results. The obtained solutions correspond to results from [6] (with provision for discretisation), but allow to avoid a computationally intense iterative procedure (see table 1 in [6]).

Let us explain in more detail fig. 1, which shows the state trajectories of the nominal system corresponding to (20) under the optimal open-loop worst-case control $u^0(t)$, $t = 0, \ldots, 119$ (dashed line *1*), and under the optimal initial open-loop control $u_0^0(t|x_0)$, $t = 0, \ldots, 79$ (solid line *2*). A dotted line represents the set $X(T|x_0, u(\cdot))$ of possible states of system (20) under the optimal open-loop worst-case control. This set lies entirely in the terminal set X_T (grey area), which illustrates that constraints (2) are satisfies with guarantees. The optimal initial open-loop control $u_0^0(\cdot|x_0)$ generates the set $X(T_1|x_0, u_0^0(\cdot|x_0))$ of possible states of system (20) at the closing instant T_1. This set belongs to the set $X_1(\alpha^0)$ (dotted lines at the bottom of fig. 1), i. e. for any $x_1 \in X(T_1|x_0, u_0^0(\cdot|x_0))$ the inequality $J_1(x_1) \ge \alpha^0$ holds. For a satisfactory representation of the set $X_1(\alpha^0)$, 83 vectors were required, i. e. $m_1 = 83$. Point x_1^* corresponds to the extremal value of the function J_1, i. e. $J_1(x_1^*) = \alpha^0$. Despite the approximation of the set $X_1(\alpha^0)$, the last equality holds exactly. The state trajectory that corresponds to the optimal open-loop worst-case control $u_1^0(\cdot|x_1^*)$ for the state $x(T_1) = x_1^*$ is shown by dash-dotted line *3*. Note that geometrically

$$J(u^0) = \min x_2,\ x \in X(T|x_0, u^0(\cdot)),\ V(\pi_1^0) = \alpha^0 = \min x_2,\ x \in X(T|x_1^*, u_1^0(\cdot|x_1^*)).$$

Figure 2 represents the optimal open-loop worst-case control $u^0(t)$, $t = 0, \ldots, 119$ (dashed line), optimal initial open-loop control $u_0^0(t|x_0)$, $t = 0, \ldots, 79$ (solid line before the closing instant), optimal open-loop worst-case control $u_1^0(t|x_1^*)$, $t = 80, \ldots, 119$ (solid line after the closing instant), and the trajectories that correspond to the worst-case disturbance. The latter here is the disturbance that delivers the exact optimal value. In the example under consideration, the worst disturbances for the optimal open-loop worst-case control and for the optimal strategy coincide and are equal to $w^*(t) = w_{\max} \operatorname{sign}(c' A^{T-t-1} M)$, $t = 0, \ldots, T-1$.

Example 2. The following problem was solved in [8]:

$$\int_0^{t_f} |u(t)| \to \min,$$

$$\dot{x}_1 = x_2,\ \dot{x}_2 = -x_1 + u + w,\ x(0) = x_0, \tag{21}$$

$$|x(t_f)| \le x^*,\ |u(t)| \le 1,\ |w(t)| \le w^*,\ t \in [0, t_f].$$

Let $u(t)=u_1(t)-u_2(t),\ 0\le u_j(t)\le 1,\ j=1,\ 2,\ x_3(t)=\int\limits_0^t u_1(s)+u_2(s)ds,\ t\in\left[0,\ t_f\right]$, and suppose that the control and the disturbance are discrete with the sampling period equal to $h=0,\ 1$. In this case we obtain the discrete problem with $n=3$, $r=2$:

$$x_4(T)\to\min,$$

$$x(t+1)=\begin{pmatrix}0.9950 & 0.0998 & 0\\ -0.0998 & 0.9950 & 0\\ 0 & 0 & 1\end{pmatrix}x(t)+\begin{pmatrix}0.0050 & -0.0050\\ 0.0998 & -0.0998\\ 0.1 & 0.1\end{pmatrix}u(t)+\begin{pmatrix}0.0005\\ 0.1051\\ 0\end{pmatrix}w(t), \tag{22}$$

$$x(0)=\bar{x}_0,\ \left|x_i(T)\right|\le x^*,\ i=1,\ 2,\ 0\le u_j(t)\le 1,\ \ j=1,\ 2,\ \left|w(t)\right|\le w^*,\ t=0,\ \ldots,\ T-1,$$

where $T=\dfrac{t_f}{h}$, $\bar{x}_0=\left(x_0,\ 0\right)$. We assume that parameters are as in [8]: $t_f=10$ resulting in $T=100$, $\bar{x}_0=\left(5,\ 0,\ 0\right)$, $x^*=2$, $w^*=0{,}3$. The closing instant $T_1=80$ is chosen.

The optimal open-loop worst-case control $u^0(\cdot)$ of problem (22) gives the optimal guaranteed value equal to $J\left(u^0\right)=5.722\,047$. The optimal control strategy has the optimal guaranteed value $V\left(\pi_1^0\right)=\alpha^0=5.039\,103$. Compared to [8], where $J\left(u^0\right)=5.656\,317$ and $V\left(\pi_1^0\right)=5.013\,186\,5$, slightly worse performance is due to discrete disturbance, while in [8] disturbance was assumed piecewise continuous. We are not presenting the optimal controls and trajectories here since they visually coincide with the results in [8].

The principal difference in solving problem (22) and applying the method from [8] to solve problem (21) is that the function $f(\alpha)$ as defined by (12) for problem (22) is linear in the parameter α, while for problem (21) this function is piecewise linear. This results in lower dimension of problem (19). The latter has T_1+1 variables, while the resulting problem in [8] has T variables. As a result, to obtain the optimal strategy by solution of problem (19) we needed only 0.081 2 s (0.75 s in [8]). The disadvantage of the method proposed in this paper compared to [8] is that we approximate the set $\Xi_1(\alpha)$ in $\mathbb{R}^3$ instead of $\mathbb{R}^2$ in [8].

Example 3. Consider a modification of problem (21)

$$\int\limits_0^{t_f}\left|u(t)\right|\to\min,$$

$$\dot{x}_1=x_2,\ \dot{x}_2=-x_1+x_3+w,\ \dot{x}_3=u,\ x(0)=x_0,$$

$$\left|x_i\left(t_f\right)\right|\le x^*,\ i=1,\ 2,\ \left|u(t)\right|\le 1,\ \left|w(t)\right|\le w^*,\ t\in\left[0,\ t_f\right].$$

Here a modification concerns the so-called indirect control of system (21).

Introducing $u(t)=u_1(t)-u_2(t),\ x_4(t)=\int\limits_0^t u_1(s)+u_2(s)ds,\ t\in\left[0,\ t_f\right]$, as in example 2, we obtain the discrete problem with $n=4$, $r=2$:

$$x_4(T)\to\min,$$

$$x(t+1)=\begin{pmatrix}0.9950 & 0.0998 & 0.0050 & 0\\ -0.0998 & 0.9950 & 0.0998 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}x(t)+\begin{pmatrix}0.0002 & -0.0002\\ 0.0050 & -0.0050\\ 0.1 & -0.1\\ 0.1 & 0.1\end{pmatrix}u(t)+\begin{pmatrix}0.0005\\ 0.1051\\ 0\\ 0\end{pmatrix}w(t), \tag{23}$$

$$x(0)=\bar{x}_0,\ \left|x_i\left(t_f\right)\right|\le x^*,\ i=1,\ 2,\ 0\le u_j(t)\le 1,\ \ j=1,\ 2,\ \left|w(t)\right|\le w^*,\ t=0,\ \ldots,\ T-1.$$

The closing instant $T_1=60$ is chosen. We illustrate the solution for the initial condition $\bar{x}_0=\left(7,\ 0,\ 0,\ 0\right)$, and the parameters $x^*=1.5$, $w^*=0.2$.

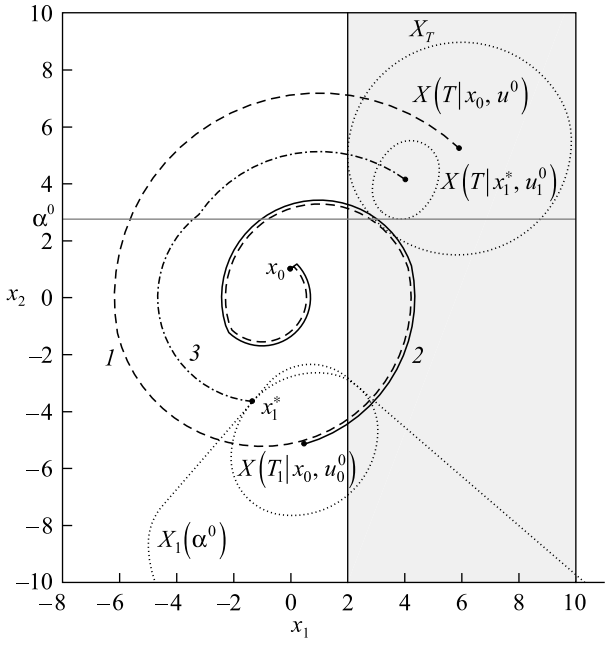

Fig. 1. Phase-plane solution representation of example 1. Trajectories under the optimal open-loop worst-case control (line *1*), under the optimal initial open-loop control (line *2*), and the optimal open-loop worst-case control on the interval after the closing instant for a sample state x_1^* (line *3*)

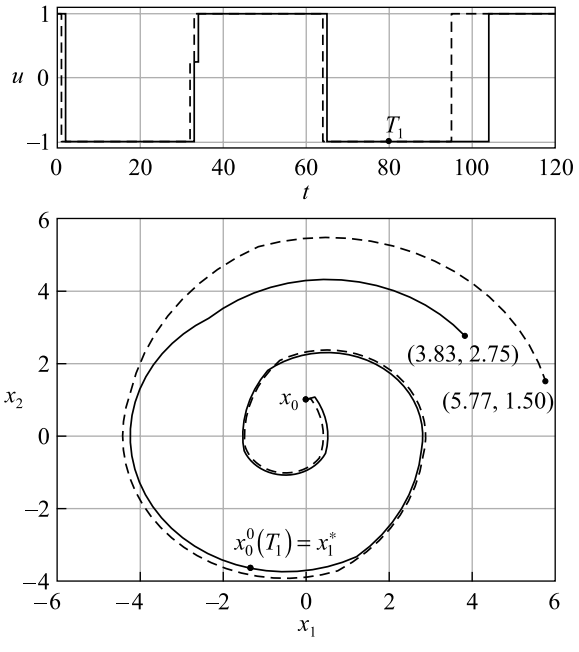

Fig. 2. Optimal control and trajectories in example 1 under the worst-case disturbance $w^*(\cdot)$

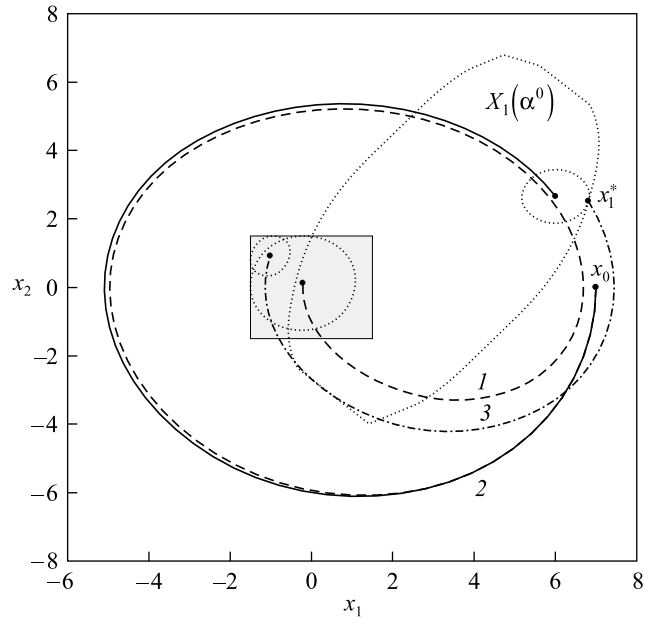

Fig. 3. Phase-plane solution representation of example 2. Trajectories under the optimal open-loop worst-case control (line *1*), under the optimal initial open-loop control (line *2*), and the optimal open-loop worst-case control on the interval after the closing instant for a sample state x_1^* (line *3*)

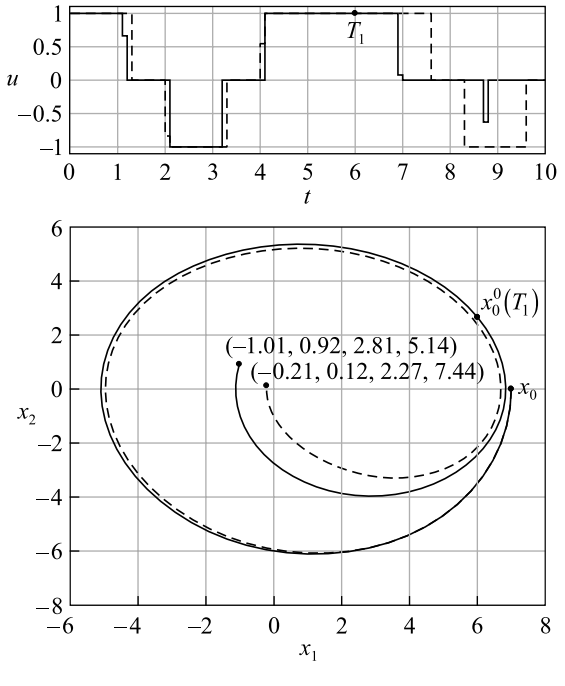

Fig. 4. Optimal control and trajectories in example 2 under the worst-case disturbance $w^*(\cdot)$

The optimal open-loop worst-case control $u^0(\cdot)$ in problem (23) has the optimal guaranteed value equal to $J\left(u^0\right) = 7.438\,263$. The optimal strategy gives $V\left(\pi_1^0\right) = \alpha^0 = 6.657\,643$. The time spent to construct the optimal open-loop worst-case control was 0.015 s, while for the optimal control strategy 0.139 s were spent.

Figures 3 and 4 show results for problem (23). In fig. 3 projections of the state trajectories on the phase plane x_1x_2, the terminal set, sets of possible states and the intersection of the set $X_1\left(\alpha^0\right)$ by a plane $\left\{x_3 = 1.966\,4,\ x_4 = 4.166\,4\right\}$, where the point $x_0^0\left(T_1\right)$ and the set $X\left(T_1 \middle| x_0,\ u_0^0\left(\cdot \middle| x_0\right)\right)$ lie, are shown. To approximate sets $X_1(\alpha)$ we used $m_1 = 1385$ vectors. In the neighbourhood of the point $x_1^* = \left(6.813\,7,\ 2.519\,2,\ 1.966\,3,\ 4.166\,4\right)$ the approximation accuracy is $J_1\left(x_1^*\right) - \alpha^0 = 6.8 \cdot 10^{-5}$.

In fig. 4 the optimal open-loop worst-case control $u^0(\cdot)$, the optimal initial open-loop control $u_0^0\left(\cdot \middle| x_0\right)$ and the realisations of the particular optimal strategy and the corresponding trajectory in the process with the disturbance defined as in example 1, are presented. The mentioned disturbance is the worst for the optimal open-loop worst-case control $u^0(\cdot)$. The optimal strategy in the same process has the value equal to 5.136 926.

Conclusion

This paper considers a guaranteed terminal cost minimisation problem for linear discrete systems with unknown bounded disturbance. We study two types of control inputs that achieve the guaranteed constraint satisfaction and minimise the cost in the problem under consideration. The first is the optimal open-loop worst-case control that is constructed entirely before the control process starts, is not corrected during the process, and ignores any possible information about the system's future behaviour. The second is the optimal control strategy with one closing instant, where closure means taking into account a state measurement at one future time instant. The optimal control strategy consists of the optimal initial open-loop control, defined at time instances before the closure, and a collection of optimal open-loop worst-case controls, defined after the closing instant for all possible (due to disturbance and initial control) states at that closing instant. Practical application of the optimal strategy implies using the optimal initial open-loop control before the closing instant and then choosing optimal open-loop worst-case control depending on the state measurement in a particular control process.

While optimal control strategies with one closing instant for linear terminal problems were introduced in [6], the main contributions of this paper consist both in the new formulation of the problem for constructing the optimal initial open-loop control and the numerical method for its solution. The proposed formulation is a minimax optimal control problem with a cost function that is implicitly defined as the optimal value of another optimal control problem. We thoroughly elaborated the structure of this problem using the duality theory, which allowed us to reduce it to an equivalent linear program and significantly simplify the method for optimal strategy construction compared to the algorithm introduced in [6]. Numerical experiments demonstrate effectiveness of the proposed approach and superiority of the optimal control strategy over the optimal open-loop worst case control.

Библиографические ссылки

1. Witsenhausen H. A minimax control problem for sampled linear systems. *IEEE Transactions on Automatic Control*. 1968;13(1): 5–21. DOI: 10.1109/TAC.1968.1098788.

2. Куржанский АБ. *Управление и наблюдение в условиях неопределенности.* Гусев МИ, редактор. Москва: Наука; 1977. 392 с.

3. Красовский НН. *Управление динамической системой. Задача о минимуме гарантированного результата.* Москва: Наука; 1985. 520 с.

4. Lee JH, Zhenghong Yu. Worst-case formulations of model predictive control for systems with bounded parameters. *Automatica*. 1997;33(5):763–781. DOI: 10.1016/S0005-1098(96)00255-5.

5. Bemporad A, Borrelli F, Morari M. Min-max control of constrained uncertain discrete-time linear systems. *IEEE Transactions on Automatic Control*. 2003;48(9):1600–1606. DOI: 10.1109/TAC.2003.816984.

6. Балашевич НВ, Габасов Р, Кириллова ФМ. Построение оптимальных обратных связей по математическим моделям с неопределенностью. *Журнал вычислительной математики и математической физики.* 2004;44(2):265–286.

7. Kostyukova O, Kostina E. Robust optimal feedback for terminal linear-quadratic control problems under disturbances. *Mathematical programming*. 2006;107(1–2):131–153. DOI: 10.1007/s10107-005-0682-4.

8. Дмитрук НМ. Оптимальная стратегия с одним моментом замыкания в линейной задаче оптимального гарантированного управления. *Журнал вычислительной математики и математической физики.* 2018;58(5):664–681. DOI: 10.7868/S0044466918050010.

9. Boyd S, Vandenberghe L. *Convex optimization*. New York: Cambridge University Press; 2004. 716 p.

10. Gal T. *Postoptimal analyses, parametric programming and related topics: degeneracy, multicriteria decision making redundancy*. Berlin: De Gruyter; 1995. 437 p.

References

1. Witsenhausen H. A minimax control problem for sampled linear systems. *IEEE Transactions on Automatic Control*. 1968;13(1): 5–21. DOI: 10.1109/TAC.1968.1098788.

2. Kurzhanskii AB. *Upravlenie i nablyudenie v usloviyakh neopredelennosti* [Control and observation under uncertainty conditions]. Gusev MI, editor. Moscow: Nauka; 1977. 392 p. Russian.

3. Krasovskii NN. *Upravlenie dinamicheskoi sistemoi. Zadacha o minimume garantirovannogo rezul'tata* [Control of a dynamical system. The problem of the minimum of the guaranteed result]. Moscow: Nauka; 1985. 520 p. Russian.

4. Lee JH, Zhenghong Yu. Worst-case formulations of model predictive control for systems with bounded parameters. *Automatica*. 1997;33(5):763–781. DOI: 10.1016/S0005-1098(96)00255-5.

5. Bemporad A, Borrelli F, Morari M. Min-max control of constrained uncertain discrete-time linear systems. *IEEE Transactions on Automatic Control*. 2003;48(9):1600–1606. DOI: 10.1109/TAC.2003.816984.

6. Balashevich NV, Gabasov R, Kirillova FM. [The construction of optimal feedback from mathematical models with uncertainty]. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*. 2004;44(2):265–286. Russian.

7. Kostyukova O, Kostina E. Robust optimal feedback for terminal linear-quadratic control problems under disturbances. *Mathematical programming*. 2006;107(1–2):131–153. DOI: 10.1007/s10107-005-0682-4.

8. Dmitruk NM. [Optimal strategy with one closing instant for a linear optimal guaranteed control problem]. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*. 2018;58(5):664–681. DOI: 10.7868/S0044466918050010. Russian.

9. Boyd S, Vandenberghe L. *Convex optimization*. New York: Cambridge University Press; 2004. 716 p.

10. Gal T. *Postoptimal analyses, parametric programming and related topics: degeneracy, multicriteria decision making redundancy*. Berlin: De Gruyter; 1995. 437 p.

Received 22.03.2021 / revised 22.06.2021 / accepted 22.06.2021.

УДК 517.925.7

О РЕШЕНИЯХ УРАВНЕНИЯ ШАЗИ

К. Г. АТРОХОВ[1)]***, Е. В. ГРОМАК***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Система Шази определяет необходимые и достаточные условия отсутствия подвижных критических точек у решений дифференциального уравнения третьего порядка, рассмотренного Шази в одной из первых работ по классификации обыкновенных дифференциальных уравнений высших порядков относительно свойства Пенлеве. Решение полной системы Шази в случае постоянных полюсов уже получено. Однако до сих пор вопрос об интегрировании уравнения Шази оставался открытым. В настоящей работе доказывается, что в случае постоянных полюсов при некоторых дополнительных условиях это уравнение интегрируется в эллиптических функциях.

Ключевые слова: уравнение Шази; система Шази; свойство Пенлеве; эллиптические функции.

ON SOLUTIONS OF THE CHAZY EQUATION

K. G. ATROKHAU[a]***, E. V. GROMAK***[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: E. V. Gromak (lenagromak@tut.by)

The Chazy system determines the necessary and sufficient conditions for the absence of movable critical points of solutions of the particular third order differential equation that was considered by Chazy in one of the first papers on the classification of higher-order ordinary differential equations with respect to the Painlevé property. The solution of the complete Chazy system in the case of constant poles has been already obtained. However, the question of integrating the Chazy equation remained open until now. In this paper, we prove that in the case of constant poles, under some additional conditions, this equation is integrated in elliptic functions.

Keywords: Chazy equation; Chazy system; Painlevé property; elliptic functions.

Образец цитирования:
Атрохов КГ, Громак ЕВ. О решениях уравнения Шази. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:51–59 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-51-59

For citation:
Atrokhau KG, Gromak EV. On solutions of the Chazy equation. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:51–59.
https://doi.org/10.33581/2520-6508-2021-2-51-59

Авторы:
Кирилл Георгиевич Атрохов – старший преподаватель кафедры дифференциальных уравнений и системного анализа механико-математического факультета.
Елена Валерьевна Громак – кандидат физико-математических наук; доцент кафедры теории функций механико-математического факультета.

Authors:
Kiryl G. Atrokhau, senior lecturer at the department of differential equations and systems analysis, faculty of mechanics and mathematics.
kiryl.atrokhau@gmail.com
https://orcid.org/0000-0002-5869-9964
Elena V. Gromak, PhD (physics and mathematics); associate professor at the department of theory of functions, faculty of mechanics and mathematics.
lenagromak@tut.by
https://orcid.org/0000-0003-3646-6227

Introduction

It is known that the Painlevé equations appeared as a result of solving the classification problem regarding the Painlevé property for the second order ordinary differential equations [1; 2]. For the equations of higher orders, the Painlevé problem, which consists in determining the conditions for the absence of movable multi-valued singularities for solutions, in the general case remains open. At present, the most complete results have been obtained only for polynomial equations of higher orders. So, for example, equations of the form

$$y''' = P\left(z, y, y', y''\right),$$

where $P\left(z, y, y', y''\right)$ is a polynomial of y and its derivatives are considered in [3–10]. Some classes of equations of the fourth and higher orders were studied in [11; 12].

One of the first works on the classification of higher-order equations with respect to the Painlevé property was a paper by Chazy [3]. It deals with the Painlevé property of the equation

$$y''' = \sum_{k=1}^{6} \frac{\left(y' - a_k'\right)\left(y'' - a_k''\right)}{y - a_k} + \sum_{k=1}^{6} \frac{A_k\left(y' - a_k'\right)^3 + B_k\left(y' - a_k'\right)^2 + C_k\left(y' - a_k'\right)}{y - a_k} + \\ + Dy'' + Ey' + \prod_{k=1}^{6}\left(y - a_k\right)\sum_{k=1}^{6}\frac{F_k}{y - a_k}, \tag{1}$$

where the poles $a_k = a_k\left(z\right)$ are finite, distinct and in general are functions of the independent variable z.

The paper [3] also contains the system of 31 algebraic and differential equations

$$(\mathcal{A}) \qquad \sum_{j=1}^{6} A_j = 0,\ \sum_{j=1}^{6} a_j A_j = -6,\ \sum_{j=1}^{6} a_j^2 A_j = -2\sum_{j=1}^{6} a_j,\ 2A_k^2 + \sum_{j=1}^{6}\frac{A_k - A_j}{a_k - a_j} = 0,\ k = 1, \ldots, 6\ (j \neq k),$$

$$(\mathcal{B}) \qquad \sum_{j=1}^{6}\left(B_k - B_j\right)\left(-\frac{A_k}{2} - \frac{1}{a_k - a_j}\right) + A_k' - \sum_{j=1}^{6}\frac{a_k' - a_j'}{a_k - a_j}\left(A_k - 3A_j\right) - \frac{3}{2}A_k\sum_{i=1}^{6} a_i' A_i = 0,$$

$$(\mathcal{C}) \qquad \left(-2A_kC_k - \sum_{j=1}^{6}\frac{C_k - C_j}{a_k - a_j}\right) + \sum_{j=1}^{6}\frac{3A_j\left(a_k' - a_j'\right)^2 + \left(2B_j - B_k\right)\left(a_k' - a_j'\right) + a_k'' - a_j''}{a_k - a_j} - \\ - B_k^2 + B_k' - B_kD + E = 0,$$

$$(\mathcal{D}) \qquad 2D + \sum_{j=1}^{6}\left(B_j - 3a_j'A_j\right) = 0,$$

$$(\mathcal{F}) \qquad \sum_{j=1}^{6} F_j = 0,\ \sum_{j=1}^{6} a_jF_j = 0,\ \sum_{j=1}^{6} a_j^2F_j = 0,\ -a_k''' - B_kC_k + C_k' + D\left(a_k'' - C_k\right) + Ea_k' + F_k\prod_{j=1}^{6}\left(a_k - a_j\right) + \\ + \sum_{j=1}^{6}\frac{A_j\left(a_k' - a_j'\right)^3 + B_j\left(a_k' - a_j'\right)^2 - \left(C_k - C_j\right)\left(a_k' - a_j'\right) + \left(a_k' - a_j'\right)\left(a_k'' - a_j''\right)}{a_k - a_j} = 0,$$

for 26 unknown functions $A_k = A_k\left(z\right), B_k = B_k\left(z\right), C_k = C_k\left(z\right), D = D\left(z\right), E = E\left(z\right), F_k = F_k\left(z\right)$. In [3, p. 367–369] Chazy claimed that the solution of this system determines necessary and sufficient conditions for the absence of movable critical points of solutions of the equation (1). In this case, the poles a_k are the parameters of the system. The equations of the system $\left(\mathcal{A}\right)-\left(\mathcal{F}\right)$ we denote below by $\left(\mathcal{A}_1\right), \ldots, \left(\mathcal{A}_9\right), \left(\mathcal{B}_1\right), \ldots, \left(\mathcal{B}_6\right), \left(\mathcal{C}_1\right), \ldots, \left(\mathcal{C}_6\right), \left(\mathcal{D}\right), \left(\mathcal{F}_1\right), \ldots, \left(\mathcal{F}_9\right)$.

The solution of the Chazy $\mathcal{A}$-system was obtained in [13] and the solution of the complete Chazy system in the case of constant poles a_k in expanded form is given in [14]. However, the question of integrating the Chazy equation (1) remained open until now. In this paper, we prove that in the case of constant poles, under some additional conditions, the equation (1) is integrated in elliptic functions.

Let us briefly summarise some results from [13; 14] which we need to obtain the main result stated in theorem 3.

Solution of system $(\mathcal{A})$

First, let us note that the successive elimination of the variables A_k allows us uniquely express A_6, A_5, A_4, A_3, A_2 in terms of A_1. Indeed from the system $(\mathcal{A})$ we successively obtain

$$A_6 = -A_5 - A_4 - A_3 - A_2 - A_1,\ A_5 = -\frac{6}{a_{56}} - \frac{a_{46}}{a_{56}}A_4 - \frac{a_{36}}{a_{56}}A_3 - \frac{a_{26}}{a_{56}}A_2 - \frac{a_{16}}{a_{56}}A_1,$$

$$A_4 = \frac{2\left(a_{41}+a_{42}+a_{43}-a_{45}-a_{46}\right)}{a_{45}a_{46}} - \frac{a_{35}a_{36}}{a_{45}a_{46}}A_3 - \frac{a_{25}a_{26}}{a_{45}a_{46}}A_2 - \frac{a_{15}a_{16}}{a_{45}a_{46}}A_1, \tag{2}$$

$$A_3 = \frac{2a_{13}\left(-a_{12}-a_{13}+2a_{14}+2a_{15}+2a_{16}\right)}{a_{34}a_{35}a_{36}} - \frac{a_{13}a_{24}a_{25}a_{26}}{a_{12}a_{34}a_{35}a_{36}}A_2 +$$

$$+ \frac{a_{13}a_{14}a_{15}a_{16}}{a_{34}a_{35}a_{36}}\left(\frac{1}{a_{12}}+\frac{1}{a_{13}}+\frac{2}{a_{14}}+\frac{2}{a_{15}}+\frac{2}{a_{16}}\right)A_1 + \frac{2a_{13}a_{14}a_{15}a_{16}}{a_{34}a_{35}a_{36}}A_1^2,$$

where a_{ij} denote the non-zero pole differences $a_i - a_j$. Let's also note the structure of A_2:

$$A_2 = \frac{n_0A_1^4 + n_1A_1^3 + n_2A_1^2 + n_3A_1 + n_4}{\left(a_2-a_4\right)\left(a_2-a_5\right)\left(a_2-a_6\right)Q_2}, \tag{3}$$

where $Q_2 = d_0A_1^2 + d_1A_1 + d_2$, and the coefficients $n_0, \ldots, n_4$ и d_0, d_1, d_2 are determined through $a_1, \ldots, a_6$.

After the substitution of the expressions thus obtained into the system $(\mathcal{A})$, the first four equations of this system become identities, and the equations $(\mathcal{A}_5)-(\mathcal{A}_9)$ acquire the form

$$\frac{a_{21}^2u_{45}u_{46}u_{56}U}{a_{24}^2a_{25}^2a_{26}^2Q_2^2} = 0,\ \frac{a_{31}^2u_{45}u_{46}u_{56}U}{a_{34}^2a_{35}^2a_{36}^2Q_2^2} = 0,\ \frac{a_{41}^2u_{56}U_4U}{a_{42}^2a_{43}^2a_{45}^2a_{46}^2Q_2^2} = 0,$$

$$\frac{a_{51}^2u_{46}U_5U}{a_{52}^2a_{53}^2a_{54}^2a_{56}^2Q_2^2} = 0,\ \frac{a_{61}^2u_{45}U_6U}{a_{62}^2a_{63}^2a_{64}^2a_{65}^2Q_2^2} = 0,$$

respectively, where $a_{ij} = a_i - a_j$, $u_{ij} = 2a_1 - a_i - a_j + \left(a_1-a_i\right)\left(a_1-a_j\right)A_1$; U and U_4, U_5, U_6 are polynomials of the fifth and second degrees in A_1, respectively.

Next two theorems from [14] follow from the successive considering of the two cases ($u_{ij} = 0$ and $U = 0$).

Lemma 1. *The system* $(\mathcal{A})$ *admits the symmetry* $\left(A_k,\ a_k\right) \leftrightarrow \left(A_j,\ a_j\right),\ j, k = 1, \ldots, 6.$

This lemma shows that the permutation of arbitrary components $\left(A_k,\ a_k\right)$ of the solution of the system $(\mathcal{A})$ with arbitrary components $\left(A_j,\ a_j\right)$ leads to the solution of the system $(\mathcal{A})$.

Theorem 1. *The system* $(\mathcal{A})$ *has the solution*

$$A_j = \frac{1}{a_5-a_j} + \frac{1}{a_6-a_j},\ j = 1, \ldots, 4,$$

$$A_5 = \frac{1}{a_1-a_5} + \frac{1}{a_2-a_5} + \frac{1}{a_3-a_5} + \frac{1}{a_4-a_5} + \frac{2}{a_5-a_6}, \tag{4}$$

$$A_6 = \frac{1}{a_1-a_6} + \frac{1}{a_2-a_6} + \frac{1}{a_3-a_6} + \frac{1}{a_4-a_6} + \frac{2}{a_6-a_5}$$

under the following condition on the poles

$$6s_4 - 3s_3a_5 + s_2a_5^2 + \left(-3s_3 + 4s_2a_5 - 3s_1a_5^2\right)a_6 + \left(s_2 - 3s_1a_5 + 6a_5^2\right)a_6^2 = 0, \tag{5}$$

where $s_1, \ldots, s_4$ *are elementary symmetric polynomials in* $a_1, \ldots, a_4$.

Consideration of the case $U = 0$ requires that the equation for A_1 has the form $p_0A_1^5 + p_1A_1^4 + p_2A_1^3 + p_3A_1^2 + p_4A_1 + p_5 = 0$, where p_i are polynomials in $a_1, \ldots, a_6$.

Theorem 2. *Let A_1 be a solution of the fifth degree equation for some fixed values of the poles a_k such that $Q_2 \neq 0$. Then A_1 and A_k ($k = 2, \ldots, 6$), evaluated on the basis of this value of A_1 and formulas (2), (3), define the solution of the system $(\mathcal{A})$.*

Thus, theorems 1 and 2 determine the solution of the Chazy $\mathcal{A}$-system. It should be noted that theorem 1 is a special case of theorem 2. However, under condition (5) the A_k can be determined in the closed form (4).

Solution of system $(\mathcal{B})-(\mathcal{F})$

The solution of the system $(\mathcal{A})-(\mathcal{F})$ for known A_k is reduced to the successive solution of three linear algebraic systems with additional constraints. In the general case, the solution of the systems $(\mathcal{B})$, $(\mathcal{C})$, $(\mathcal{F})$ can be obtained by the Gauss method. Therefore, the system $(\mathcal{B})$ with known A_k is a linear system $A_B B = R_B$ with respect to $B = (B_1, \ldots, B_6)^T$, where matrix A_B has the entries:

$$\{A_B\}_{kj} = \frac{A_k}{2} + \frac{1}{a_k - a_j},\ j \neq k,\ \{A_B\}_{kk} = -\frac{5A_k}{2} - \sum_{i=1}^{6} \frac{1}{a_k - a_i},\ i \neq k.$$

In general, the rank of this matrix does not exceed five. The vector R_B depends only on A_k, a_k and their derivatives. Under the condition of theorem 1, the matrix A_B can be represented in closed form as well. To do this we need to substitute the values $A_1, \ldots, A_6$ from (4) into the above matrix A_B.

With known A_k and B_k from the equation $(\mathcal{D})$ we find that

$$D = -\frac{1}{2}\sum_{j=1}^{6} B_j + \frac{3}{2} a_j' A_j.$$

The system $(\mathcal{C})$ with respect to $C = (C_1, \ldots, C_6)^T$ is also linear and can be represented in the form $A_C C = R_C$ with the matrix A_C:

$$\{A_C\}_{kj} = \frac{1}{a_k - a_j},\ j \neq k;\ \{A_C\}_{kk} = -2A_k - \sum_{i=1}^{6} \frac{1}{a_k - a_i},\ i \neq k.$$

The rank of the matrix A_C also does not exceed five. In this case, the inhomogeneity vector R_C contains A_k, B_k, a_k, their derivatives and the unknown function $E(z)$. Under the condition of theorem 1, the matrix A_C can also be written explicitly. To do this we substitute the values $A_1, \ldots, A_6$ from (4) into the above matrix A_C.

The system $(\mathcal{F}_4)-(\mathcal{F}_9)$ is also linear with respect to $F = (F_1, \ldots, F_6)^T$. In this case, the inhomogeneity matrix R_F contains the unknown function $E(z)$. However, the substitution of F_k into the equations $(\mathcal{F}_1)-(\mathcal{F}_3)$ allows us to define $E(z)$.

A case of constant poles

The Chazy equation in the case of constant poles a_k has the form

$$y''' = y'y''\sum_{k=1}^{6}\frac{1}{y-a_k} + y'^3\sum_{k=1}^{6}\frac{A_k}{y-a_k} + y'^2\sum_{k=1}^{6}\frac{B_k}{y-a_k} + \\ + y'\sum_{k=1}^{6}\frac{C_k}{y-a_k} + Dy'' + Ey' + \prod_{k=1}^{6}(y-a_k)\sum_{k=1}^{6}\frac{F_k}{y-a_k}. \tag{6}$$

The system $(\mathcal{A})$ remains the same, and the system $(\mathcal{B})-(\mathcal{F})$ is greatly simplified and acquires the form

$$(\mathcal{B}^*) \qquad \left(-\frac{5}{2}A_k - \sum_{j=1}^{6}\frac{1}{a_k - a_j}\right)B_k + \sum_{j=1}^{6}\left(\frac{A_k}{2} + \frac{1}{a_k - a_j}\right)B_j = 0,\ k = 1, \ldots, 6 \quad (j \neq k),$$

$$(\mathcal{C}^*) \qquad \left(-2A_k - \sum_{j=1}^{6}\frac{1}{a_k - a_j}\right)C_k + \sum_{j=1}^{6}\frac{C_j}{a_k - a_j} - B_k^2 + B_k' - B_k D + E = 0,$$

$$(\mathcal{D}^*) \qquad 2D + \sum_{j=1}^{6} B_j = 0,$$

$$(\mathcal{F}^*) \qquad \sum_{j=1}^{6} F_j = 0,\ \sum_{j=1}^{6} a_j F_j = 0,\quad \sum_{j=1}^{6} a_j^2 F_j = 0,\ -B_k C_k - DC_k + C_k' + F_k \prod_{j=1}^{6} \left(a_k - a_j\right) = 0.$$

The solution of the $\mathcal{A}$-system in closed form is given by theorem 1. The solution of the system $\left(\mathcal{B}^*\right)-\left(\mathcal{F}^*\right)$ in closed form is obtained in [14]. Using these results, we investigate the integrability of the Chazy equation (6).

Let us consider the case when $B_k = 0$, $k = 1, \ldots, 6$. The following statements are true regarding to the definition of the remaining coefficients of the Chazy equation (6).

Lemma 2. *If in the Chazy equation (6) A_k are determined by theorem 1 and $B_k = 0$, $k = 1, \ldots, 6$, then the fulfillment of the Chazy system $\left(\mathcal{B}^*\right)-\left(\mathcal{F}^*\right)$ necessarily leads to*

$$D = 0,\ F_k = 0,\ k = 1, \ldots, 6.$$

Lemma 3. *Let the following conditions be fulfilled in the Chazy equation (6):*

A1. *The coefficients A_k and the poles a_k are defined by theorem 1 and B_k, C_k, D, E satisfy the Chazy system $\left(\mathcal{B}^*\right)-\left(\mathcal{F}^*\right)$.*

A2. $B_k = 0$, $k = 1, \ldots, 6$.

A3. *The constant poles a_k satisfy the condition*

$$h_1 := a_1 + a_2 + a_3 + a_4 - a_5 - 3a_6 \neq 0.$$

Then $C_6 = \mathcal{C}_6$, $E = \mathcal{E}$, where $\mathcal{C}_6$, $\mathcal{E}$ are the arbitrary constants and $C_1 - C_5$ are determined by the formulas

$$C_j = \frac{\mathcal{E}}{h_1}\left(a_j - a_6\right)^2 + \mathcal{C}_6 \frac{\left(2a_j - a_5 - a_6\right)h_1 - 2\left(a_j - a_6\right)^2}{\left(a_5 - a_6\right)h_1},\ j = 1, \ldots, 4,$$

$$C_5 = -\frac{\mathcal{E}}{h_1}\left(a_5 - a_6\right)^2 - \mathcal{C}_6 \frac{h_1 - 2a_5 + 2a_6}{h_1}.$$

Lemma 4. *If the conditions* **A1**, **A2** *of lemma 3 are fulfilled and $h_1 = 0$, then $C_5 = \mathcal{C}_5$, $E = \mathcal{E}$, where $\mathcal{C}_5$, $\mathcal{E}$ are the arbitrary constants and $C_1 - C_4$, C_6 are determined by the formulas*

$$C_j = -\mathcal{E}\frac{\left(a_j - a_5\right)^2}{2\left(a_5 - a_6\right)} - \mathcal{C}_5 \frac{\left(a_j - a_6\right)^2}{\left(a_5 - a_6\right)^2},\ j = 1, \ldots, 4,$$

$$C_6 = \frac{\mathcal{E}\left(a_5 - a_6\right)}{2}.$$

The proof of lemmas 2–4 follows directly from the result of the paper [14], where the solution of the Chazy system (6) is given in the case of theorem 1. The above formulas for determining the remaining coefficients of the equation (6) follow from the corresponding formulas in the paper [14] with $B_k = 0$, $k = 1, \ldots, 6$.

The general integral of the equation (6) will be sought in the form

$$\left(w'\right)^2 = K_1 P(w) + K_2 Q(w) + R(w), \tag{7}$$

where K_1, K_2 are the arbitrary coefficients and $P(w)$, $Q(w)$, $R(w)$ are the polynomials of w with constant coefficients not higher than the fourth degree. In this case, the third constant of integration is obtained by separating the variables in the equation (7) and its integration.

It is clear that in the case of the existence of such polynomials $P(w)$, $Q(w)$, $R(w)$, the solution of (7) and, therefore, the solution of the equation (6) is generally expressed in a rational way in terms of the Weierstrass elliptic function $\wp(z)$.

Differentiating twice (7) and excluding arbitrary constants K_1 and K_2, we find

$$w''' = w'w''\frac{P''Q - PQ''}{P'Q - PQ'} - w'^3\frac{P''Q' - P'Q''}{2\left(P'Q - PQ'\right)} + w'\frac{Q''\left(P'R - PR'\right) - P''\left(R'Q - RQ'\right)}{2\left(P'Q - PQ'\right)}.$$

Hereinafter the primes on the polynomials $P(w)$, $Q(w)$, $R(w)$ denote the corresponding derivatives with respect to w. Comparing this equation with the equation (6) we have three conditions for the definition of the polynomials $P(w)$, $Q(w)$, $R(w)$:

$$\frac{P''Q - PQ''}{P'Q - PQ'} = \sum_{k=1}^{6} \frac{1}{w - a_k}, \tag{8}$$

$$-\frac{P''Q' - P'Q''}{2\left(P'Q - PQ'\right)} = \sum_{k=1}^{6} \frac{A_k}{w - a_k}, \tag{9}$$

$$\frac{P\left(Q'R'' - Q''R'\right) + Q\left(R'P'' - R''P'\right) + R\left(P'Q'' - P''Q'\right)}{2\left(P'Q - PQ'\right)} = E + \sum_{k=1}^{6} \frac{C_k}{w - a_k}, \tag{10}$$

moreover $B_k = 0$, $F_k = 0$, $D = 0$.

Without loss of generality in (8)–(10) we consider

$$P'Q - PQ' = \prod_{k=1}^{6} \left(w - a_k\right), \tag{11}$$

where $P = \sum_{j=0}^{4} p_j w^j$, $Q = \sum_{j=0}^{4} q_j w^j$, $R = \sum_{j=0}^{4} r_j w^j$ and in this case

$$p_4 = 0,\ q_4 = 1,\ q_3 = 0. \tag{12}$$

Then from the condition (8) which has the form (11) we find

$$\begin{gathered} p_3 = -1,\ p_2 = \frac{\sigma_1}{2},\ p_1 = \frac{1}{3}\left(-q_2 - \sigma_2\right), \\ p_0 = \frac{1}{4}\left(-2q_1 + \sigma_3\right),\ q_0 = \frac{1}{18}\left(2q_2^2 + 3q_1\sigma_1 + 2q_2\sigma_2 - 6\sigma_4\right), \end{gathered} \tag{13}$$

where σ_k here and below are the symmetric polynomials with respect to $a_1, \ldots, a_6$, and q_1, q_2 must satisfy the following two conditions:

$$\begin{gathered} -\frac{1}{4}q_1\left(-2q_1 + \sigma_3\right) - \frac{1}{54}\left(q_2 + \sigma_2\right)\left(2q_2^2 + 3q_1\sigma_1 + 2q_2\sigma_2 - 6\sigma_4\right) - \sigma_6 = 0, \\ q_1q_2 - \frac{q_2q_3}{2} + \frac{1}{18}\sigma_1\left(2q_2^2 + 3q_1\sigma_1 + 2q_2\sigma_2 - 6\sigma_4\right) + \sigma_5 = 0. \end{gathered} \tag{14}$$

The condition (9) recorded by the virtue of (11) in the form

$$P''Q' - P'Q'' = -2\left(P'Q - PQ'\right)\sum_{k=1}^{6} \frac{A_k}{w - a_k},$$

under the fulfillment of (12), (13) and A_k from theorem 1, defines q_1, q_2:

$$\begin{gathered} q_1 = \frac{2}{3}\left(a_3a_4 + a_2\left(a_3 + a_4\right) + a_1\left(a_2 + a_3 + a_4\right)\right)\left(a_5 + a_6\right), \\ q_2 = -\left(a_1 + a_2 + a_3 + a_4\right)a_5 - \left(a_1 + a_2 + a_3 + a_4 - 2a_5\right)a_6. \end{gathered} \tag{15}$$

Note that the values q_1, q_2, defined by (15), satisfy the conditions (14).

The condition (10) by the virtue of (11) takes the form

$$P\left(Q'R'' - Q''R'\right) + Q\left(R'P'' - R''P'\right) + R\left(P'Q'' - P''Q'\right) = 2\left(P'Q - PQ'\right)\left(E + \sum_{k=1}^{6} \frac{C_k}{w - a_k}\right). \tag{16}$$

This condition leads to the determination of the coefficients of the polynomial R. Wherein two cases must be considered corresponding to lemmas 3 and 4.

In the case of lemma 3, that is $h_1 \neq 0$, the coefficients of the polynomial R are determined by the relations

$$\begin{aligned} r_0 = -\Big(&\left(2\mathcal{C}_6 - \mathcal{E}a_5\right)\left(-3a_2a_3a_4 - 3a_1\left(a_3a_4 + a_2\left(a_3 + a_4\right)\right) + \left(a_3a_4 + a_2\left(a_3 + a_4\right) + a_1\left(a_2 + a_3 + a_4\right)\right)a_5\right) + \\ &+ \left(-3\mathcal{E}\left(a_2a_3a_4 + a_1\left(a_2a_3 + \left(a_2 + a_3\right)a_4\right)\right) + 3\mathcal{E}\left(a_1 + a_2 + a_3 + a_4\right)a_5^2 + \right. \\ &\left. + 2\mathcal{C}\left(a_2a_3 + a_2a_4 + a_3a_4 + 3\left(a_2 + a_3 + a_4\right)a_5 - 6a_5^2 + a_1\left(a_2 + a_3 + a_4 + 3a_5\right)\right)\right)a_6 + \end{aligned}$$

$$+\left(\mathcal{E}\left(a_3a_4+a_2\left(a_3+a_4\right)+a_1\left(a_2+a_3+a_4\right)\right)-3\left(4\mathcal{C}+\mathcal{E}\left(a_1+a_2+a_3+a_4\right)\right)a_5-\right.$$
$$\left.-12\mathcal{E}a_5^2\right)a_6^2+12\mathcal{E}a_5a_6^3\Big)/\left(6h_1\left(a_5-a_6\right)\right),$$
$$r_1=\Big(2\Big(\mathcal{E}\left(a_5-a_6\right)\left(a_3a_4+a_2\left(a_3+a_4\right)+a_1\left(a_2+a_3+a_4\right)-3a_6\left(a_5+a_6\right)\right)+\mathcal{C}\Big(-2a_3a_4+3a_3a_5+$$
$$+3a_4a_5-3a_5^2+3\left(a_3+a_4-2a_5\right)a_6-3a_6^2+a_2\left(-2a_3-2a_4+3\left(a_5+a_6\right)\right)+$$
$$+a_1\left(-2a_2-2a_3-2a_4+3\left(a_5+a_6\right)\right)\Big)\Big)\Big)/\left(3h_1\left(a_5-a_6\right)\right), \tag{17}$$
$$r_2=-\mathcal{E},\ r_3=0,\ r_4=0,$$

where $C_6=\mathcal{C}_6$, $E=\mathcal{E}$ and $\mathcal{C}_6$, $\mathcal{E}$ are the arbitrary constants. The rest of the equations of the condition (16) become identity due to (17) and the values of the coefficients C_k from lemma 3 and the condition of theorem 1.

In the case of lemma 4, that is $h_1=0$, the coefficients of the polynomial R are determined by the relations

$$r_0=\Big(a_4\left(a_4-a_5\right)a_5\left(2\mathcal{C}_5+\mathcal{E}a_5\right)+\Big(2\mathcal{C}_5a_4^2+3a_5^2\left(2\mathcal{C}_5-3\mathcal{E}a_5\right)-a_4a_5\left(8\mathcal{C}_5+3\mathcal{E}a_5\right)\Big)a_6+\Big(-a_4\left(6\mathcal{C}_5+\mathcal{E}a_4\right)+$$
$$+\left(-6\mathcal{C}_5+\mathcal{E}a_4\right)a_5+18\mathcal{E}a_5^2\Big)a_6^2+3\mathcal{E}\left(a_4-3a_5\right)a_6^3-a_3^2\left(3a_4-a_5-a_6\right)\left(2\mathcal{C}_5+\mathcal{E}a_5-\mathcal{E}a_6\right)-$$
$$-a_3\left(a_4-a_5-3a_6\right)\left(3a_4-a_5-a_6\right)\left(2\mathcal{C}_5+\mathcal{E}a_5-\mathcal{E}a_6\right)-a_1^2\left(3a_3+3a_4-a_5-a_6\right)\left(2\mathcal{C}_5+\mathcal{E}a_5-\mathcal{E}a_6\right)-$$
$$-a_1\left(a_3+a_4-a_5-3a_6\right)\left(3a_3+3a_4-a_5-a_6\right)\Big(2\mathcal{C}_5+\mathcal{E}a_5-\mathcal{E}a_6\Big)\Big)/\Big(12\left(a_5-a_6\right)^2\Big),$$
$$r_1=\Big(2\mathcal{C}_5a_1^2+2\mathcal{C}_5a_2^2+2\mathcal{C}_5a_3^2+2\mathcal{C}_5a_4^2+\mathcal{E}a_1^2a_5+\mathcal{E}a_2^2a_5+\mathcal{E}a_3^2a_5+\mathcal{E}a_4^2a_5-2\mathcal{C}_5a_5^2+5\mathcal{E}a_5^3- \tag{18}$$
$$-\Big(4\mathcal{C}_5a_1+\mathcal{E}a_1^2+4\mathcal{C}_5a_2+\mathcal{E}a_2^2+4\mathcal{C}_5a_3+\mathcal{E}a_3^2+4\mathcal{C}_5a_4+\mathcal{E}a_4^2+$$
$$+2\left(-2\mathcal{C}_5+\mathcal{E}\left(a_1+a_2+a_3+a_4\right)\right)a_5+3\mathcal{E}a_5^2\Big)a_6+$$
$$+\Big(6\mathcal{C}_5+2\mathcal{E}\left(a_1+a_2+a_3+a_4\right)-5\mathcal{E}a_5\Big)a_6^2+3\mathcal{E}a_6^3\Big)/\Big(6\left(a_5-a_6\right)^2\Big),$$
$$r_2=-\mathcal{E},\ r_3=0,\ r_4=0,$$

where $C_5=\mathcal{C}_5$, $E=\mathcal{E}$ and $\mathcal{C}_5$, $\mathcal{E}$ are the arbitrary constants. The rest of the equations of (16) become identity due to (18) and the values of the coefficients C_k from lemma 4 and the condition of theorem 1. The above considerations imply the following statement.

Theorem 3. *If in the Chazy equation (6) $B_k=0$, $k=1,\ldots,6$, then under the conditions of the Chazy system $\left(\mathcal{B}^*\right)-\left(\mathcal{F}^*\right)$ and A_k, defined by theorem 1, the equation (6) is generally integrated in elliptic functions.*

P r o o f. If $B_k=0$ and $\left(\mathcal{B}^*\right)-\left(\mathcal{F}^*\right)$ hold then by lemma 1 $D=0$ and $F_k=0$. By the virtue of theorem 1, we have two cases: $h_1\neq 0$ and $h_1=0$ to define C_k. In both cases, the polynomials $P(w)$, $Q(w)$, $R(w)$ are chosen according to the formulas (12), (13), (15), (17), (18) which proves the statement of the theorem.

Note that lemmas 3 and 4, on which the proof of theorem 3 is based, are given in [15; 16], respectively.

Now consider the following example.

Let $a_2=1$, $a_3=-1$, $a_4=2$, $a_5=-2$, $a_6=0$. Then from (5) we find $a_1=-\frac{8}{5}$, and from (4) we obtain the solution of the system $(\mathcal{A})$:

$$A_1=-\frac{15}{8},\ A_2=-\frac{4}{3},\ A_3=0,\ A_4=-\frac{3}{4},\ A_5=\frac{37}{12},\ A_6=\frac{7}{8}.$$

In this case $h_1=\frac{12}{5}$. Therefore, setting $B_k=0$, $k=1,\ldots,6$, from the Chazy system we have $D=0$ and $F_k=0$, and from lemma 3

$$C_1=\frac{5}{3}\mathcal{C}_6+\frac{16}{15}\mathcal{E},\ C_2=\frac{1}{12}\left(-19\mathcal{C}_6+5\mathcal{E}\right),\ C_3=\frac{5}{12}\left(\mathcal{C}_6+\mathcal{E}\right),\ C_4=\frac{1}{3}\left(-4\mathcal{C}_6+5\mathcal{E}\right),$$
$$C_5=-\frac{1}{3}\left(8\mathcal{C}_6+5\mathcal{E}\right),\ C_6=\mathcal{C}_6,\ E=\mathcal{E}.$$

The coefficients $C_6 = \mathcal{C}_6$, $E = \mathcal{E}$ are remain arbitrary, and the coefficients of the polynomials $P(w)$ and $Q(w)$, respectively, have the following form:

$$p_4 = 0,\ p_3 = -1,\ p_2 = -\frac{4}{5},\ p_1 = \frac{7}{5},\ p_0 = -\frac{4}{5},$$

$$q_4 = 1,\ q_3 = 0,\ q_2 = \frac{4}{5},\ q_1 = \frac{28}{5},\ q_0 = -\frac{16}{5}.$$

From (17) we find the coefficients of the polynomial $R(w)$:

$$r_4 = r_3 = 0,\ r_2 = -\mathcal{E},\ r_1 = \frac{1}{6}\left(5\mathcal{C}_6 - 7\mathcal{E}\right),\ r_0 = \frac{2}{3}\left(\mathcal{C}_6 + \mathcal{E}\right).$$

The general integral of the Chazy equation in this case has the form

$$\left(w'\right)^2 = b_0 + b_1 w + b_2 w^2 + b_3 w^3 + b_4 w^4, \tag{19}$$

where

$$b_0 = \frac{2}{15}\left(6K_1 + 24K_2 - 5\mathcal{C}_6 - 5\mathcal{E}\right),\ b_1 = -\frac{7}{5}K_1 - \frac{28}{5}K_2 - \frac{5}{6}\mathcal{C}_6 + \frac{7}{6}\mathcal{E},$$

$$b_2 = \frac{4}{5}\left(K_1 - K_2\right) + \mathcal{E},\ b_3 = K_1,\ b_4 = -K_2$$

and K_1, K_2 are the arbitrary constants. The third arbitrary constant appears from the separation of variables in the equation (19) and its integration. Thus, for example, if $K_2 = 0$, $K_1 \neq 0$, then

$$w = \alpha\rho(z) + \beta,$$

where $\alpha = \frac{4}{K_1}$, $\beta = -\frac{\mathcal{E}}{3K_1} - \frac{4}{15}$ and $\rho(z)$ is the elliptic Weierstrass function satisfying the equation

$$\left(\rho'\right)^2 = 4\rho^3 - g_2\rho - g_3,$$

$$g_2 = \frac{242K_1^2 + 125K_1\mathcal{C}_6 - 95K_1\mathcal{E} + 50\mathcal{E}^2}{600},$$

$$g_3 = \frac{-8176K_1^3 - 500\mathcal{E}^3 + 75K_1\mathcal{E}\left(19\mathcal{E} - 25\mathcal{C}_6\right) + 30K_1^2\left(100\mathcal{C}_6 + 83\mathcal{E}\right)}{108\ 000}.$$

Библиографические ссылки

1. Ince EL. *Ordinary Differential Equations*. New York, Dover, 1956. 558 p.
2. Iwasaki K, Kimura H, Shimomura S, Yoshida M. *From Gauss to Painlevé: a Modern Theory of Special Functions*. Braunschweig: Vieweg; 1991. 347 p. (Aspects of mathematics; volume 16).
3. Chazy J. Sur les équations différentielles du troisième ordre et d'ordre supérieur dont l'intégrale générale a ses points critiques fixes. *Acta Mathematica.* 1911;34(1):317–385.
4. Bureau FJ. Differential equations with fixed critical points. *Annali di Matematica Pura ed Applicata.* 1964;64(1):229–364. DOI: 10.1007/BF02410054.
5. Bureau FJ. Differential equations with fixed critical points. *Annali di Matematica Pura ed Applicata.* 1964;66(1):1–116. DOI: 10.1007/BF02412437.
6. Martynov IP. Third-order equations without moving critical singularities. *Differential Equations.* 1985;21(6):937–946.
7. Exton H. Nonlinear ordinary differential equations with fixed critical points. *Rendiconti di Matematica.* 1973;6(2):419–462.
8. Cosgrove CM. Higher-order Painlevé equations in the polynomial class I. Bureau symbol P2. *Studies in Applied Mathematics.* 2000;104(1):1–65. DOI: 10.1111/1467-9590.00130.
9. Mŭgan U, Jrad F. Painlevé test and higher order differential equations. *Journal of Nonlinear Mathematical Physics*. 2002;9(3): 282–310. DOI: 10.2991/jnmp.2002.9.3.4.
10. Cosgrove CM. Higher-order Painlevé equations in the polynomial class II. Bureau symbol P1. *Studies in Applied Mathematics.* 2006;116(4):321–413. DOI: 10.1111/j.1467-9590.2006.00346.x.
11. Kudryashov NA. Fourth-order analogies to the Painlevé equations. *Journal of Physics A: Mathematical and General.* 2002; 35(21):4617–4632. DOI: 10.1088/0305-4470/35/21/310.
12. Sobolevsky S. Painlevé classification of binomial type ordinary differential equations of the arbitrary order. *Studies in Applied Mathematics.* 2006;117(3):215–237. DOI: 10.1111/j.1467-9590.2006.00353.x.
13. Gromak VI. On solutions of the Chazy system. *Differential Equations.* 2007;43(5):631–635. DOI: 10.1134/S0012266107050060.
14. Atrokhov KG, Gromak VI. Solution of the Chazy system. *Differential Equations.* 2010;46(6):783–797. DOI: 10.1134/S0012266110060030.

15. Громак ЕВ. Об интегрировании уравнения Шази с постоянными полюсами в эллиптических функциях. В: Рогозин СВ, редактор. *Тезисы докладов международной научной конференции «Аналитические методы анализа и дифференциальных уравнений»; 11–14 сентября 2012 г.; Минск, Беларусь.* Минск: Институт математики Национальной академии наук Беларуси; 2012. с. 28.

16. Громак ЕВ. Об уравнениях третьего порядка Р-типа. В: Деменчук АК, Красовский СГ, Макаров ЕК, редакторы. *XVI Международная научная конференция по дифференциальным уравнениям (Еругинские чтения – 2014); 20–22 мая 2014 г.; Новополоцк, Беларусь. Часть 1.* Минск: Институт математики НАН Беларуси; 2014. с. 11–12.

References

1. Ince EL. *Ordinary Differential Equations*. New York, Dover, 1956. 558 p.

2. Iwasaki K, Kimura H, Shimomura S, Yoshida M. *From Gauss to Painlevé: a Modern Theory of Special Functions*. Braunschweig: Vieweg; 1991. 347 p. (Aspects of mathematics; volume 16).

3. Chazy J. Sur les équations différentielles du troisième ordre et d'ordre supérieur dont l'intégrale générale a ses points critiques fixes. *Acta Mathematica.* 1911;34(1):317–385.

4. Bureau FJ. Differential equations with fixed critical points. *Annali di Matematica Pura ed Applicata.* 1964;64(1):229–364. DOI: 10.1007/BF02410054.

5. Bureau FJ. Differential equations with fixed critical points. *Annali di Matematica Pura ed Applicata.* 1964;66(1):1–116. DOI: 10.1007/BF02412437.

6. Martynov IP. Third-order equations without moving critical singularities. *Differential Equations.* 1985;21(6):937–946.

7. Exton H. Nonlinear ordinary differential equations with fixed critical points. *Rendiconti di Matematica.* 1973;6(2):419–462.

8. Cosgrove CM. Higher-order Painlevé equations in the polynomial class I. Bureau symbol P2. *Studies in Applied Mathematics.* 2000;104(1):1–65. DOI: 10.1111/1467-9590.00130.

9. Mǔgan U, Jrad F. Painlevé test and higher order differential equations. *Journal of Nonlinear Mathematical Physics*. 2002;9(3): 282–310. DOI: 10.2991/jnmp.2002.9.3.4.

10. Cosgrove CM. Higher-order Painlevé equations in the polynomial class II. Bureau symbol P1. *Studies in Applied Mathematics.* 2006;116(4):321–413. DOI: 10.1111/j.1467-9590.2006.00346.x.

11. Kudryashov NA. Fourth-order analogies to the Painlevé equations. *Journal of Physics A: Mathematical and General.* 2002; 35(21):4617–4632. DOI: 10.1088/0305-4470/35/21/310.

12. Sobolevsky S. Painlevé classification of binomial type ordinary differential equations of the arbitrary order. *Studies in Applied Mathematics.* 2006;117(3):215–237. DOI: 10.1111/j.1467-9590.2006.00353.x.

13. Gromak VI. On solutions of the Chazy system. *Differential Equations.* 2007;43(5):631–635. DOI: 10.1134/S0012266107050060.

14. Atrokhov KG, Gromak VI. Solution of the Chazy system. *Differential Equations.* 2010;46(6):783–797. DOI: 10.1134/S0012266110060030.

15. Gromak EV. [On integration of the Chazy equation with constant poles in elliptic functions]. In: Rogozin SV, editor. *Tezisy dokladov mezhdunarodnoi nauchnoi konferentsii «Analiticheskie metody analiza i differentsial'nykh uravnenii»; 11–14 sentyabrya 2012 g.; Minsk, Belarus'* [Abstracts of the International scientific conference «Analytical methods of analysis and differential equations»; 2012 September 11–14; Minsk, Belarus]. Minsk: Institute of Mathematics, Natioanl Academy of Sciences of Belarus; 2012. p. 28. Russian.

16. Gromak EV. [On the third-order P-type equations]. In: Demenchuk AK, Krasovskii SG, Makarov EK, editors. *XVI Mezhdunarodnaya nauchnaya konferentsiya po differentsial'nym uravneniyam (Eruginskie chteniya – 2014); 20–22 maya 2014 g.; Novopolotsk, Belarus'. Chast' 1* [XVI International scientific conference on differential equations (Erugin readings – 2014); 2014 May 20–22; Novopolotsk, Belarus. Part 1]. Minsk: Institute of Mathematics, National Academy of Sciences of Belarus; 2014. p. 11–12. Russian.

Received 28.04.2021 / revised 01.07.2021 / accepted 01.07.2021.

Теория вероятностей и математическая статистика

Probability theory and mathematical statistics

УДК 519.2

СТАТИСТИЧЕСКАЯ ПОСЛЕДОВАТЕЛЬНАЯ ПРОВЕРКА ГИПОТЕЗ О ПАРАМЕТРАХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ СЛУЧАЙНЫХ БИНАРНЫХ ДАННЫХ

А. Ю. ХАРИН[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается актуальная математическая задача компьютерного анализа данных – задача статистической последовательной проверки простых гипотез о параметрах распределения вероятностей наблюдаемых бинарных данных. Эта задача решается для двух моделей наблюдений: схемы независимых испытаний и однородной цепи Маркова. Выведены легко интерпретируемые и удобные для компьютерной реализации явные выражения статистик последовательных тестов (статистических критериев). Разработан подход для вычисления характеристик эффективности решающих правил – вероятностей ошибочных решений и математических ожиданий случайного числа наблюдений, необходимых для обеспечения требуемой точности. Получены асимптотические разложения для указанных характеристик эффективности при «засорениях» распределения вероятностей наблюдаемых данных.

Ключевые слова: случайные бинарные данные; простые гипотезы; статистический последовательный тест; вероятность ошибки; математическое ожидание случайного числа наблюдений; «засорения»; асимптотические разложения.

Благодарность. Работа выполнена при финансовой поддержке Министерства образования Республики Беларусь в рамках государственной программы научных исследований «Конвергенция-2025».

Образец цитирования:
Харин АЮ. Статистическая последовательная проверка гипотез о параметрах распределений вероятностей случайных бинарных данных. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:60–66 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-60-66

For citation:
Kharin AYu. Statistical sequential hypotheses testing on parameters of probability distributions of random binary data. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:60–66.
https://doi.org/10.33581/2520-6508-2021-2-60-66

Автор:
Алексей Юрьевич Харин – доктор физико-математических наук, доцент; заведующий кафедрой теории вероятностей и математической статистики факультета прикладной математики и информатики.

Author:
Alexey Yu. Kharin, doctor of science (physics and mathematics), docent; head of the department of probability theory and mathematical statistics, faculty of applied mathematics and computer science.
kharinay@bsu.by

STATISTICAL SEQUENTIAL HYPOTHESES TESTING ON PARAMETERS OF PROBABILITY DISTRIBUTIONS OF RANDOM BINARY DATA

A. Yu. KHARIN[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

An important mathematical problem of computer data analysis – the problem of statistical sequential testing of simple hypotheses on parameters of probability distributions of observed binary data – is considered in the paper. This problem is being solved for two models of observation: for independent observations and for homogeneous Markov chains. Explicit expressions of the sequential tests statistics are derived, transparent for interpretation and convenient for computer realisation. An approach is developed to calculate the performance characteristics – error probabilities and mathematical expectations of the random number of observations required to guarantee the requested accuracy for decision rules. Asymptotic expansions for the mentioned performance characteristics are constructed under «contamination» of the probability distributions of observed data.

Keywords: random binary data; simple hypotheses; statistical sequential test; error probability; mathematical expectation of the random number of observations; «contamination»; asymptotic expansions.

Acknowledgements. This work was supported by Ministry of Education of the Republic of Belarus within the state research program «Convergence-2025».

Introduction

Data become one of the active drivers of the world economy, and computer data analysis becomes an essential part of the modern life. Efficiency of data analysis defines the success in a growing spectrum of fields. Binary data is a very important class of data for several reasons: 1) binary data is natural for computer processing; 2) binary data describes many situations in terms of «presence – absence», «positive – negative», etc.; 3) binary data can be used for description of a significantly rich family of observations, if considered by groups. Classical methods of statistical analysis are often not applied to such data, as those methods assumptions are usually not satisfied for the binary data models, or they are not effective.

As deterministic approach has a limited potential to describe the processes, real-life data is usually considered to be random, and probabilistic models are used. In these models, an important problem that often appears, is a problem of discrimination between two typical situations on the probability distributions of random data. These two typical situations can be formulated in terms of simple hypotheses on parameters of the probability distribution, and the problem turns to the problem of statistical testing of two simple hypotheses [1].

In many cases, especially in statistical quality control, in automatic warning systems, in personalised medicine, in financial decision making, it is important to use the minimal number of observations that guarantee the requested accuracy [2]. Sequential statistical tests [3] follow this principle with the assumption that the number of observations to be used is defined through the observation process, depend on observations themselves, and thus is a random variable. Due to the complex structure of sequential decision rules, usually theoretical analysis of their performance characteristics – error probabilities and mathematical expectations of the random number of required observations – is problematic [4].

In the paper we develop an approach to calculate and analyse the performance characteristics of sequential tests for binary data. In practice the factual probability distribution of data often deviates from the hypothetical one – the hypothetical probability distribution is «contaminated» [5; 6], so we also consider this situation here, and analyse deviations of the performance characteristics under distortion.

Results for the model of independent binary random observations

Denote by $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}$ correspondently sets of: positive integer, integer, and real numbers; $\mathbf{Z}_+ = \mathbf{N} \cup \{0\}$. Let independent identically distributed K-dimensional binary random vectors

$$x_t = (x_{ti}) \in U = \left\{u^1, \ldots, u^{2^K}\right\} = \left\{\begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix}, u_i \in B = \{0, 1\}, i = 1, \ldots, K\right\}, t \in \mathbf{N}, \quad (1)$$

be observed in a probability space (Ω, F, P) with independent components. The probability distribution of random vectors (1) depends on the unknown value of the parameters vector $\theta = \begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix} \in \Theta = \{\theta_0, \theta_1\}$, where $\theta_0 = \begin{pmatrix} p_1^0 \\ \vdots \\ p_K^0 \end{pmatrix}$, $\theta_1 = \begin{pmatrix} p_1^1 \\ \vdots \\ p_K^1 \end{pmatrix}$, $\theta_0 \neq \theta_1$. Such a model is often used in practice to decide in favour of two possible alternative typical situations. The components of the vector of parameters $p_i \in \{p_i^0, p_i^1\}$, $p_i^0, p_i^1 \in (0, 1)$, mean the probabilities of the random event $\{x_{ti} = 1\}$, $t \in \mathbf{N}$, $i = 1, \ldots, K$.

Suppose the following assumption is satisfied for the probability distribution of binary random vectors (1):

$$P(u; \theta) = P_\theta\{x_t = u\} = a^{-J(u;\theta)}, \ t \in \mathbf{N}, \ u \in U, \tag{2}$$

where $a \in \mathbf{R}$, $a > 1$; with $J(u; \theta)$: $U \times \Theta \to \mathbf{Z}_+$ being a function that satisfies the following condition:

$$\sum_{u \in U} a^{-J(u;\theta)} = 1. \tag{3}$$

There are two simple hypotheses considered on the parameters vector value θ of probability distribution (2):

$$H_0: \theta = \theta_0, \ H_1: \theta = \theta_1. \tag{4}$$

Denote the accumulated log-likelihood ratio test statistic:

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{t=1}^{n} \lambda_t, \tag{5}$$

where

$$\lambda_t = \lambda(x_t) = \log_a \left(\frac{P(x_t; \theta_1)}{P(x_t; \theta_0)} \right) \tag{6}$$

is the log-likelihood ratio for the binary observation vector x_t.

Theorem 1. *For the model (1)–(3) the sequence of statistics (5) for hypotheses (4) is a homogeneous Markov chain with discrete time, and it has the form*

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{k=1}^{2^K} n_k \left(J(u^k; \theta_0) - J(u^k; \theta_1) \right) \in \mathbf{Z}, \ n \in \mathbf{N}, \tag{7}$$

where n_k denotes the number of the vectors that equal u^k observed within n binary random vectors $\{x_1, \ldots, x_n\}$, $\sum_{k=1}^{2^K} n_k = n$.

P r o o f. The Markov property [7] of the random sequence (5) follows from the independence of its increments (6) due to the independence of random observations (1). Expression (7) is derived by equivalent transformations of (5), (6) under the assumption (2), (3).

Corollary 1. *If under theorem 1 conditions $K = 1$, then test statistic (5) is*

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = n_0 \left(J(0; p^0) - J(0; p^1) \right) + (n - n_0) \left(J(1; p^0) - J(1; p^1) \right), \ n \in \mathbf{N},$$

where n_0 denotes the number of the observations equal to 0.

Using statistics (7), the sequential probability ratio test [1] for hypotheses (4) is constructed as follows: the decision after n observations made ($n = 1, 2, \ldots$) is

$$d = d(n) = \mathbf{1}_{[C_+, +\infty)}(\Lambda_n) + 2 \cdot \mathbf{1}_{(C_-, C_+)}(\Lambda_n), \tag{8}$$

where $\mathbf{1}_D(\cdot)$ denotes the indicator function of a set D. Decisions $d = 0$ and $d = 1$ mean the observation process termination and acceptance of correspondently H_0 or H_1 after n binary random vectors observed; $d = 2$ means that the $(n + 1)$ vector should be observed; $C_-, C_+ \in \mathbf{Z}$, $C_- < C_+$ are parameters of the decision rule (8) called thresholds; in practice they are often calculated according to [3]:

$$C_- = \left[\log_a\left(\frac{\beta_0}{1-\alpha_0}\right)\right],\ C_+ = \left[\log_a\left(\frac{1-\beta_0}{\alpha_0}\right)\right], \tag{9}$$

where α_0, β_0 are the admissible values of error type I (to accept H_1, when H_0 is true) and error type II (H_1 is true, H_0 is accepted) probabilities; $[\cdot]$ means the integer part of an argument. With thresholds (9) the factual values of error probabilities of type I and II may differ from α_0, β_0, and the problem of the factual values calculation of the performance characteristics is open for the sequential tests.

Introduce the notation: $\delta_{i,j}$ for the Kronecker delta; $\mathbf{I}_k$ for the identity matrix of size k; $\mathbf{0}_{m\times n}$ for the zero-matrix of size $(m \times n)$; $\mathbf{1}(\cdot)$ for the unit step function; $\mathbf{1}_k$ for the k-vector column with components equal 1. Denote by $t^{(k)}$ the expected value of the random number of observations (sample size) provided the true hypothesis is H_k, $k \in \{0, 1\}$, and by α, β the factual values of error type I and II probabilities for test (8); $N = C_+ - C_-$; let $P^{(k)} = \left(p_{ij}^{(k)}\right) = \begin{pmatrix} \mathbf{I}_2 & \vdots & \mathbf{0}_{2\times N} \\ --- & \vdots & --- \\ R^{(k)} & \vdots & Q^{(k)} \end{pmatrix}$ be the matrix of size $(N+2)\times(N+2)$, with blocks $R^{(k)}$, $Q^{(k)}$ defined by

$$p_{ij}^{(k)} = \begin{cases} \sum\limits_{u\in U} \delta_{J(u;\theta_0)-J(u;\theta_1),\,j-i} P(u;\theta_k),\ i, j \in (C_-, C_+), \\ \sum\limits_{u\in U} \mathbf{1}\left(C_- - i + J(u;\theta_1) - J(u;\theta_0)\right) P(u;\theta_k),\ i \in (C_-, C_+),\ j = C_-, \\ \sum\limits_{u\in U} \mathbf{1}\left(J(u;\theta_0 - J(u;\theta_1)) + i - C_+\right) P(u;\theta_k),\ i \in (C_-, C_+),\ j = C_+. \end{cases}$$

Denote

$$\pi^{(k)} = \left(\pi_i^{(k)}\right),\ \pi_i^{(k)} = \sum_{u\in U} \delta_{J(u;\theta_0)-J(u;\theta_1),\,i} P(u;\theta_k),\ i \in \{C_- + 1, \ldots, C_+ - 1\},$$

$$\pi_{C_+}^{(k)} = \sum_{i\ge C_+}\sum_{u\in U} \delta_{J(u;\theta_0)-J(u;\theta_1),\,i} P(u;\theta_k),\ \pi_{C_-}^{(k)} = \sum_{i\le C_-}\sum_{u\in U} \delta_{J(u;\theta_0)-J(u;\theta_1),\,i} P(u;\theta_k),$$

$$S^{(k)} = \mathbf{I}_N - Q^{(k)},\ B^{(k)} = \left(S^{(k)}\right)^{-1} R^{(k)},$$

let $W_{(i)}$ means the i column of the matrix W.

Theorem 2. *If under the model (1)–(6), $\left|S^{(k)}\right| \ne 0$, $k \in \{0, 1\}$, then the performance characteristics of sequential decision rule (8) are calculated in the explicit form:*

$$t^{(k)} = \left(\pi^{(k)}\right)' \left(S^{(k)}\right)^{-1} \mathbf{1}_N + 1,\ \alpha = \left(\pi^{(0)}\right)' B_{(2)}^{(0)} + \pi_{C_+}^{(0)},\ \beta = \left(\pi^{(1)}\right)' B_{(1)}^{(1)} + \pi_{C_-}^{(1)}.$$

P r o o f is based on the theory of finite homogeneous Markov chains with discrete time. The sequence

$$\zeta_n = C_- \cdot \mathbf{1}_{(-\infty, C_-]}(\Lambda_n) + C_+ \cdot \mathbf{1}_{[C_+, +\infty)}(\Lambda_n) + \Lambda_n \cdot \mathbf{1}_{(C_-, C_+)}(\Lambda_n)$$

is a homogeneous Markov chain with $N + 2$ states, and C_-, C_+ are the absorbing states.

The situation, where the probability distribution of data is «contaminated» is considered in the next section for a more general model, where the observed binary data form a Markov chain instead of being independent.

Results for the model of observations forming a homogeneous binary Markov chain

Suppose binary random vectors $x_1, x_2, \ldots$ forming a homogeneous Markov chain are being observed, taking values in the set $U = \left\{u^1, \ldots, u^{2^K}\right\} = \left\{\begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix},\ u_i \in B = \{0, 1\},\ i = 1, \ldots, K\right\}$. To simplify notation, introduce the

set $V = \{0, 1, \ldots, M-1\}$, $M = 2^K$, one-to-one corresponding to the set U. Denote for the observed Markov chain the initial probabilities vector $\pi = (\pi_i)$, $i \in V$, and the one-step transition probabilities matrix $P = (p_{ij})$, $i, j \in V$:

$$P\{x_1 = i\} = \pi_i, \ P\{x_n = j | x_{n-1} = i\} = p_{ij}, \ i, j \in V.$$

As in the case of independent binary random vectors, consider two simple alternative hypotheses on the parameter values of the Markov chain:

$$H_0\colon \pi = \pi^{(0)}, \ P = P^{(0)}; \ H_1\colon \pi = \pi^{(1)}, \ P = P^{(1)}, \tag{10}$$

where $\pi^{(0)} = \left(\pi_i^{(0)}\right)$, $\pi^{(1)} = \left(\pi_i^{(1)}\right)$ are two given values for initial probabilities vector, $P^{(0)} = \left(p_{ij}^{(0)}\right) \neq P^{(1)} = \left(p_{ij}^{(1)}\right)$ are given matrices of one-step transition probabilities for the correspondent hypotheses.

For construction of the sequential test for this model of data, denote

$$\lambda_1 = \log \frac{\pi_{x_1}^{(1)}}{\pi_{x_1}^{(0)}}, \ \lambda_k = \log \frac{p_{x_{k-1}, x_k}^{(1)}}{p_{x_{k-1}, x_k}^{(0)}}, \ k > 1, \ \Lambda_n = \sum_{k=1}^{n} \lambda_k, \ n \in \mathbf{N}. \tag{11}$$

The sequential test for the considered model and hypotheses (10) is constructed according to the decision rule (8), with replacing (5)–(7) by (11). According to this test, for defined thresholds (see, e. g., (9)) $C_-, C_+ \in \mathbf{R}$, $C_- < 0$, $C_+ > 0$, hypothesis H_0 is accepted after n observations, if $\Lambda_n \le C_-$, hypothesis H_1 is accepted, if $\Lambda_n \ge C_+$, the test is stopped in both those cases, otherwise the test is proceeded, and the $(n+1)$ binary random vector should be observed. The sequence of $(K+1)$ dimensional random vectors $(\Lambda_n, x_n)'$, $n \in \mathbf{N}$, is a Markov chain by the definition:

$$P\{\Lambda_n, x_n | \Lambda_{n-1}, \Lambda_{n-2}, \ldots, \Lambda_1, x_{n-1}, x_{n-2}, \ldots, x_1\} = P\{\Lambda_n, x_n | \Lambda_{n-1}, x_{n-1}\}.$$

Suppose $\pi^{(0)}$, $P^{(0)}$, $\pi^{(1)}$, $P^{(1)}$ be satisfying the following assumption:

$$\exists a \in \mathbf{R}, \ m_i, m_{ij} \in \mathbf{Z}, \ i, j \in V\colon \log \frac{\pi_i^{(1)}}{\pi_i^{(0)}} = m_i a, \ \log \frac{p_{ij}^{(1)}}{p_{ij}^{(0)}} = m_{ij} a. \tag{12}$$

Without loss of generality, suppose for test (5), (8), (11) the thresholds $C_-, C_+ \in \mathbf{Z}$, and denote that $t^{(k)}$ is the expected sample size till one of the hypotheses is accepted, provided H_k is true, $k \in \{0, 1\}$;

$$W^{(k)} = \left(w_{ij}^{(k)}\right) = \begin{pmatrix} \mathbf{I}_2 & \vdots & \mathbf{0}_{2 \times MN} \\ --- & \vdots & --- \\ R^{(k)} & \vdots & Q^{(k)} \end{pmatrix}$$

is the matrix of the size $(MN+2)(MN+2)$, with blocks $R^{(k)}$, $Q^{(k)}$ defined by their elements ($s, t \in V$):

$$w_{Mi+s\, Mj+t}^{(k)} = \delta_{m_{st}, j-i} p_{st}^{(k)}, \ i, j \in (C_-, C_+),$$

$$w_{Mi+s\, Mj+t}^{(k)} = \begin{cases} \sum\limits_{t \in V} \mathbf{1}(i + m_{st} - C_+), \ j = C_+, \ i \in (C_-, C_+), \\ \sum\limits_{t \in V} \mathbf{1}(C_- - i - m_{st}), \ j = C_-, \ i \in (C_-, C_+); \end{cases} \tag{13}$$

as in the case of independent observations, denote the matrices

$$S^{(k)} = \mathbf{I}_{MN} - Q^{(k)}, \ B^{(k)} = \left(S^{(k)}\right)^{-1} R^{(k)}, \ k \in \{0, 1\};$$

the vectors

$$\omega^{(k)} = \left(\omega_i^{(k)}\right), \ i \in \{MC_- + 1, \ldots, MC_+ - 1\}\colon \omega_{Mi+s}^{(k)} = \delta_{m_s, i} \pi_s^{(k)}, \ i \in (C_-, C_+), \tag{14}$$

and the two prior probabilities of absorption

$$\omega_{MC_-}^{(k)} = \sum_{s \in V} \mathbf{1}(C_- - m_s) \pi_s^{(k)}, \ \omega_{MC_+}^{(k)} = \sum_{s \in V} \mathbf{1}(m_s - C_+) \pi_s^{(k)}. \tag{15}$$

Theorem 3. *For the observation model of a homogeneous binary vector Markov chain described above, if* $\left|S^{(k)}\right| \neq 0,\ k \in \{0, 1\}$, *and (12) holds, then the performance characteristics of sequential test (5), (8), (11) are calculated in the explicit form:*

$$t^{(k)} = \left(\omega^{(k)}\right)' \left(S^{(k)}\right)^{-1} \mathbf{1}_{MN} + 1,\ \alpha = \left(\omega^{(0)}\right)' B^{(0)}_{(2)} + \omega^{(0)}_{MC_+},\ \beta = \left(\omega^{(1)}\right)' B^{(1)}_{(1)} + \omega^{(1)}_{MC_-}.$$

P r o o f. Introduce the random sequence

$$\xi_n = MC_- \cdot \mathbf{1}_{(-\infty,\, C_-]}\left(\frac{\Lambda_n}{a}\right) + MC_+ \cdot \mathbf{1}_{[C_+,\, +\infty)}\left(\frac{\Lambda_n}{a}\right) + \left(\frac{\Lambda_n}{a} M + x_n\right) \cdot \mathbf{1}_{(C_-,\, C_+)}\left(\frac{\Lambda_n}{a}\right),\ n \in \mathbf{N},$$

that is a homogeneous Markov chain with $MN + 2$ states; two of them ($\xi_n = MC_-$ and $\xi_n = MC_+$) are absorbing. The one-step transition probabilities matrix is defined by (13), the vector of transient states initial probabilities is (14), and the initial probabilities of absorbing states are calculated in (15).

Consider now the situation that is often in practice, when the hypothetical model described above is «contaminated» in terms of the probability distribution of observations via «contamination» of the initial probabilities vector and of the one-step transition probabilities matrix. Suppose instead of the hypothetical values, the factual (distorted) vector of the initial probabilities and the factual one-step transition probabilities matrix are

$$\bar{\pi}^{(k)} = (1 - \varepsilon)\pi^{(k)} + \varepsilon\tilde{\pi}^{(k)},\ \bar{P}^{(k)} = (1 - \varepsilon)P^{(k)} + \varepsilon\tilde{P}^{(k)},\ k = 0, 1, \tag{16}$$

where $\tilde{\pi}^{(k)}$ and $\tilde{P}^{(k)}$ are the initial probabilities vector and the one-step transition probabilities matrix for the «contaminating» Markov chain, $P^{(k)} \neq \tilde{P}^{(k)}$, $k = 0, 1$, and $\varepsilon \in \left[0, \frac{1}{2}\right)$ is the probability of «contamination» (also called «contamination» level).

Denote for the «contaminated» model (16): $\tilde{W}^{(k)}$, $\tilde{Q}^{(k)}$, $\tilde{R}^{(k)}$, $\tilde{\omega}^{(k)}$, $\tilde{\omega}^{(k)}_{MC_\pm}$, $k = 0, 1$, analogously to the hypothetical case, replacing the hypothetical probability distribution by the «contaminating» one. Denote that $\tilde{S}^{(k)} = \mathbf{I}_{MN} - Q^{(k)} - \varepsilon\left(\tilde{Q}^{(k)} - Q^{(k)}\right)$.

Theorem 4. *If the hypothetical model of binary vector homogeneous Markov chain observations described above is distorted according to (16), and assumption (12) is satisfied also for the distorted model, then the error type I and II probabilities* $\bar{\alpha}$, $\bar{\beta}$, *and the conditional mathematical expectations of the observation number* $\bar{t}^{(k)}$, $k = 0, 1$, *deviate from the hypothetical performance characteristics by the values of order* $O(\varepsilon)$:

$$\bar{\alpha} - \alpha = \varepsilon\left(\left(\omega^{(0)}\right)'\left(\left(S^{(0)}\right)^{-1}\left(\left(\tilde{Q}^{(0)} - Q^{(0)}\right)\left(S^{(0)}\right)^{-1} R^{(0)} + \tilde{R}^{(0)} - R^{(0)}\right)\right)_{(2)} + \right.$$
$$\left. + \left(\tilde{\omega}^{(0)} - \omega^{(0)}\right)' B^{(0)}_{(2)} + \tilde{\omega}^{(0)}_{MC_+} - \omega^{(0)}_{MC_+}\right) + O\left(\varepsilon^2\right),$$

$$\bar{\beta} - \beta = \varepsilon\left(\left(\omega^{(1)}\right)'\left(\left(S^{(1)}\right)^{-1}\left(\left(\tilde{Q}^{(1)} - Q^{(1)}\right)\left(S^{(1)}\right)^{-1} R^{(1)} + \tilde{R}^{(1)} - R^{(1)}\right)\right)_{(1)} + \right.$$
$$\left. + \left(\tilde{\omega}^{(1)} - \omega^{(1)}\right)' B^{(1)}_{(1)} + \tilde{\omega}^{(1)}_{MC_-} - \omega^{(1)}_{MC_-}\right) + O\left(\varepsilon^2\right),$$

$$\bar{t}^{(k)} - t^{(k)} = \varepsilon\left(\left(\tilde{\omega}^{(k)} - \omega^{(k)}\right)' + \left(\omega^{(k)}\right)'\left(S^{(k)}\right)^{-1}\left(\tilde{Q}^{(k)} - Q^{(k)}\right)\right)\left(S^{(k)}\right)^{-1}\mathbf{1}_{MN} + O\left(\varepsilon^2\right).$$

P r o o f. Under «contamination» (16) the initial probabilities vector and the one-step transition probabilities matrix of the random sequence ξ_n have the correspondent mixture form, and the rest of the proof is derived by equivalent transformations.

Conclusion

An approach to calculate and analyse the performance characteristics of sequential tests for binary data for two models is proposed: for the independent binary vectors, and for the homogeneous binary vector Markov chains. The situation, where the factual model of data deviates from the hypothetical assumptions, is considered

in the paper, and correspondent differences in performance characteristics of the sequential tests are analysed asymptotically with respect to the «contamination» level. The results can be applied to construct robust sequential tests for binary random data [8], for the model of random sequences with a trend [9]. The results are also potentially applicable to the case of more than two hypotheses [10], complex hypotheses [11] and to the analysis of truncated sequential tests [12].

Библиографические ссылки

1. Mukhopadhyay N, de Silva B. *Sequential methods and their applications.* Boca Raton: CRC Press; 2009. 409 p.
2. Lai TL. *Sequential analysis: some classical problems and new challenges. Statistica Sinica.* 2001;11:303–408.
3. Wald A. Sequential analysis. New York: John Wiley and Sons; 1947. 212 p.
4. Айвазян СА. Сравнение оптимальных свойств критериев Неймана – Пирсона и Вальда. *Теория вероятностей и ее применения.* 1959;4(1):86–93.
5. Huber PJ, Ronchetti EM. *Robust statistics.* New York: Wiley; 2009. 354 p.
6. Maevskii VV, Kharin YuS. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control.* 2002;63(11):1803–1820. DOI: 10.1023/A:1020959432568.
7. Kemeny JG, Snell JL. *Finite Markov Chains.* New York: D. Van Nostrand Co.; 1960. 210 p.
8. Харин АЮ. *Робастность байесовских и последовательных статистических решающих правил.* Минск: БГУ; 2013. 207 с.
9. Kharin A, Tu TT. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics.* 2017;46(3–4):23–36. DOI: 10.17713/ajs.v46i3-4.668.
10. Tu TT, Kharin AYu. Sequential probability ratio test for many simple hypotheses on parameters of time series with trend. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;1:35–45. DOI: 10.33581/2520-6508-2019-1-35-45.
11. Kharin AY. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences.* 2017;227(2):196–203. DOI: 10.1007/s10958-017-3585-z.
12. Харин АЮ, Ту ТТ. О вычислении вероятностей ошибок усеченного последовательного критерия отношения вероятностей. *Журнал Белорусского государственного университета. Математика. Информатика.* 2018;2018(1):68–76.

References

1. Mukhopadhyay N, de Silva B. *Sequential methods and their applications.* Boca Raton: CRC Press; 2009. 409 p.
2. Lai TL. *Sequential analysis: some classical problems and new challenges. Statistica Sinica.* 2001;11:303–408.
3. Wald A. Sequential analysis. New York: John Wiley and Sons; 1947. 212 p.
4. Aivazian SA. Comparison of optimal properties of the tests of Neyman – Pearson and Wald. *Teoriya veroyatnostei i ee primeneniya.* 1959;4(1):86–93. Russian.
5. Huber PJ, Ronchetti EM. *Robust statistics.* New York: Wiley; 2009. 354 p.
6. Maevskii VV, Kharin YuS. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control.* 2002;63(11):1803–1820. DOI: 10.1023/A:1020959432568.
7. Kemeny JG, Snell JL. *Finite Markov Chains.* New York: D. Van Nostrand Co.; 1960. 210 p.
8. Kharin AY. *Robastnost' baiesovskikh i posledovatel'nykh statisticheskikh reshayushchikh pravil* [Robustness of Bayesian and sequential statistical decision rules]. Minsk: Belarusian State University; 2013. 207 p. Russian.
9. Kharin A, Tu TT. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics.* 2017;46(3–4):23–36. DOI: 10.17713/ajs.v46i3-4.668.
10. Tu TT, Kharin AY. Sequential probability ratio test for many simple hypotheses on parameters of time series with trend. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;1:35–45. DOI: 10.33581/2520-6508-2019-1-35-45.
11. Kharin AY. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences.* 2017;227(2):196–203. DOI: 10.1007/s10958-017-3585-z.
12. Kharin AY, Tu TT. On error probability calculation for the truncated sequential probability ratio test. *Journal of the Belarusian State University. Mathematics and Informatics.* 2018;2018(1):68–76. Russian.

Received 25.05.2021 / revised 28.06.2021 / accepted 28.06.2021.

Дискретная математика и математическая кибернетика

Discrete mathematics and mathematical cybernetics

УДК 681.32

РАСКРАСКА СМЕШАННОГО ГРАФА КАК ПОСТРОЕНИЕ РАСПИСАНИЯ ОБСЛУЖИВАНИЯ МНОГОПРОЦЕССОРНЫХ ТРЕБОВАНИЙ С ОДИНАКОВЫМИ ДЛИТЕЛЬНОСТЯМИ

Ю. Н. СОТСКОВ[1)]

[1)]*Объединенный институт проблем информатики НАН Беларуси, ул. Сурганова, 6, 220012, г. Минск, Беларусь*

Задача обслуживания частично упорядоченных единичных требований последовательными приборами формулируется как раскраска смешанного графа, т. е. как назначение целых чисел (цветов) $\{1, 2, \ldots, t\}$ вершинам (требованиям) $V = \{v_1, v_2, \ldots, v_n\}$ смешанного графа $G = (V, A, E)$, при котором вершины v_p и v_q, инцидентные ребру $[v_p, v_q] \in E$, имеют различные цвета. А при наличии дуги $(v_i, v_j) \in A$ цвет вершины v_i не превосходит цвет вершины v_j. Доказано, что оптимальная раскраска смешанного графа $G = (V, A, E)$ эквивалентна задаче $GcMPT \mid p_i = 1 \mid C_{\max}$ поиска оптимального расписания обслуживания частично упорядоченных требований с единичными (одинаковыми) длительностями. В отличие от классических задач построения расписаний в рассматриваемой задаче $GcMPT \mid p_i = 1 \mid C_{\max}$ необходимо несколько различных приборов для обслуживания отдельного требования. Помимо отношений предшествования, заданных на множестве требований $V = \{v_1, v_2, \ldots, v_n\}$,

Образец цитирования:
Сотсков ЮН. Раскраска смешанного графа как построение расписания обслуживания многопроцессорных требований с одинаковыми длительностями. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:67–81 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-67-81

For citation:
Sotskov YuN. Mixed graph colouring as scheduling multi-processor tasks with equal processing times. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:67–81.
https://doi.org/10.33581/2520-6508-2021-2-67-81

Автор:
Юрий Назарович Сотсков – доктор физико-математических наук, профессор; главный научный сотрудник лаборатории математической кибернетики.

Author:
Yuri N. Sotskov, doctor of science (physics and mathematics), full professor; chief researcher at the laboratory of mathematical cybernetics.
sotskov48@mail.ru
https://orcid.org/0000-0002-9971-6169

должно выполняться некоторое подмножество требований одновременно. На основании доказанных в статье теорем утверждается, что множество аналитических результатов, полученных ранее для задач $GcMPT\left|p_i=1\right|C_{\max}$, имеют аналоги для оптимальных раскрасок смешанных графов $G=(V, A, E)$, и наоборот.

Ключевые слова: оптимизация; расписание с единичными длительностями; быстродействие; смешанный граф; вершинная раскраска.

Благодарность. Это исследование выполнено при частичной финансовой поддержке Белорусского республиканского фонда фундаментальных исследований (проект № Ф21-010).

MIXED GRAPH COLOURING AS SCHEDULING MULTI-PROCESSOR TASKS WITH EQUAL PROCESSING TIMES

Yu. N. SOTSKOV [a]

[a]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, 6 Surhanava Street, Minsk 220012, Belarus*

A problem of scheduling partially ordered unit-time tasks processed on dedicated machines is formulated as a mixed graph colouring problem, i. e., as an assignment of integers (colours) $\{1, 2, \ldots, t\}$ to the vertices (tasks) $V=\{v_1, v_2, \ldots, v_n\}$ of the mixed graph $G=(V, A, E)$ such that if vertices v_p and v_q are joined by an edge $[v_p, v_q]\in E$, their colours have to be different. Further, if two vertices v_i and v_j are joined by an arc $(v_i, v_j)\in A$, the colour of vertex v_i has to be no greater than the colour of vertex v_j. We prove that an optimal colouring of a mixed graph $G=(V, A, E)$ is equivalent to the scheduling problem $GcMPT\left|p_i=1\right|C_{\max}$ of finding an optimal schedule for partially ordered multi-processor tasks with unit (equal) processing times. Contrary to classical shop-scheduling problems, several dedicated machines are required to process an individual task in the scheduling problem $GcMPT\left|p_i=1\right|C_{\max}$. Moreover, along with precedence constraints given on the set $V=\{v_1, v_2, \ldots, v_n\}$, it is required that a subset of tasks must be processed simultaneously. Due to the theorems proved in this article, most analytical results that have been proved for the scheduling problems $GcMPT\left|p_i=1\right|C_{\max}$ so far, have analogous results for optimal colourings of the mixed graphs $G=(V, A, E)$, and vice versa.

Keywords: optimisation; unit-time scheduling; makespan; mixed graph; vertex colouring.

Acknowledgements. This work was partially supported by the Belarusian Republican Foundation for Fundamental Research (project No. Ф21-010).

Introduction

Scheduling models with the prerequisite of equal (or unit) processing times to all given tasks are an approximation of coping with mass-industrial productions and manufacturing of similar items, particularly for a job-shop manufacturing problem that allows managers to personalise each individual item [1]. Such a scheduling problem with unit-times and the minimisation of the makespan is equivalent to an optimal graph colouring that consists of assigning a minimal number of colours to vertices of the graph such that no two adjacent vertices have the same colour. When a scheduling problem requires both precedence and incompatibility constraints, one needs to use a mixed graph colouring introduced in [2] for a formulation of the unit-time scheduling problem. Since the publication of article [2] in 1976, many studies of unit-time scheduling problems with the makespan criterion are based on mixed graph colourings.

Let $G=(V, A, E)$ denote a finite mixed graph with non-empty set $V=\{v_1, v_2, \ldots, v_n\}$ of the vertices placed at the first position in parenthesis, arc set A at the second position, and edge set E at the third position. An arc $(v_i, v_j)\in A$ defines the ordered pair of vertices v_i and v_j. An edge $[v_p, v_q]\in E$ means an unordered pair of vertices v_p and v_q. In what follows, we assume that a mixed graph $G=(V, A, E)$ contains no multiple arcs, no multiple edges, and no loops. If the set A is empty, we have a graph $G=(V, \varnothing, E)$. If the set E is empty, we have a digraph $G=(V, A, \varnothing)$. In article [2], a mixed graph colouring is introduced as follows.

Definition 1 [2]. An integer-valued function $c: V\to\{1, 2, \ldots, t\}$ is a colouring $c(G)$ of the mixed graph $G=(V, A, E)$, if the non-strict inequality $c(v_i)\le c(v_j)$ holds for each arc $(v_i, v_j)\in A$ and $c(v_p)\ne c(v_q)$ for

each edge $\left[v_p, v_q\right] \in E$. A mixed graph colouring $c(G)$ is optimal, if it uses a minimal possible number $\chi(G)$ of different colours $c(v_i) \in \{1, 2, \ldots, t\}$, such a minimal number $\chi(G)$ being called a chromatic number of the mixed graph $G = (V, A, E)$.

If $A = \varnothing$, a colouring $c(G)$ is the usual colouring of the vertices of the graph $G = (V, \varnothing, E)$. Contrary to a colouring of the vertices of the graph $G = (V, \varnothing, E)$ existing for any graph $G = (V, \varnothing, E)$, a mixed graph $G = (V, A, E)$ with $A \neq \varnothing$ and $E \neq \varnothing$ may be uncolourable. A criterion for the existence of a colouring $c(G)$ for the mixed graph G is proved in [2].

Theorem 1 [2]. *A colouring $c(G)$ of the mixed graph $G = (V, A, E)$ exists if and only if the digraph $(V, A, \varnothing)$ has no circuit containing adjacent vertices in the graph $(V, \varnothing, E)$.*

A mixed graph $G = (V, A, E)$ is colourable, if there exists a colouring $c(G)$ of the mixed graph G, otherwise, a mixed graph $G = (V, A, E)$ is uncolourable.

Finding an optimal colouring $c(G)$ of the mixed graph $G = (V, A, E)$ is an NP-hard problem, even if $A = \varnothing$ [3]. In articles [4; 5], it is shown that a job-shop scheduling problem with unit processing times of all operations and the minimisation of a schedule length (makespan) may be represented as an optimal colouring $c(G)$ of the specified mixed graph $G = (V, A, E)$. In article [6], it is shown that any job-shop scheduling problem with unit processing times of all operations and the minimisation of a total completion time (TCT) may be represented as a mixed graph colouring $c(G)$ minimising a sum of colours of path-endpoints of the specified mixed graph $G = (V, A, E)$ (see also articles [7; 8]).

The unit-time scheduling problem with minimising makespan is NP-hard even for three dedicated machines (processors) [9]. The complexity of a job-shop scheduling problem with a fixed number of jobs (and a fixed number of machines) is investigated in articles [10–13].

Since the NP-hard unit-time flow-shop scheduling problem [14] is polynomially reduced to the job-shop scheduling problem to minimise the TCT, the latter problem is also NP-hard. The complexity of a job-shop scheduling problem with any regular criterion is investigated in [10; 11; 13; 15]. The complexity of a mixed shop-scheduling problem is studied in [16; 17]. A different connection between mixed graph colourings and unit-time shop-scheduling problems is studied in [18–24]. Article [25] presents a comprehensive survey on mixed graph colourings and the equivalent unit-time shop-scheduling problems.

In our article, we show that an optimal colouring $c(G)$ of the mixed graph $G = (V, A, E)$ is equivalent to finding an optimal schedule for partially ordered multi-processor tasks with unit processing times (or with equal processing times). Contrary to a classical shop-scheduling problem, several dedicated machines are used simultaneously by a task during the complete processing period. Along with the precedence constraints, which are given on the set $V = \{v_1, v_2, \ldots, v_n\}$ of multi-processor tasks, it is required that a subset of tasks must be processed simultaneously. Due to the proven equivalence of the above scheduling problem and the mixed graph colouring $c(G)$, most claims that have been proved so far for a wide class of scheduling problems (without operation preemptions) have analogous claims for optimal mixed graph colourings $c(G)$, and vice versa. Throughout this article, we use the terminology from [26; 27] for graph theory and that from [28; 29] for scheduling theory.

Two classes of shop-scheduling problems as mixed graph colourings

To classify shop-scheduling problems, one can use a three-field notation $\alpha|\beta|\gamma$ introduced in [30], where α specifies a task system and machine environments, β is job characteristics, and γ is an objective function (see [29] for the extensions of classifying parameters).

General shop-scheduling problems with unit-time tasks and minimising makespan. In the general shop unit-time minimum-length scheduling problem denoted by $G|t_i = 1|C_{\max}$, a job set $J = \{J_1, J_2, \ldots, J_{|J|}\}$ must be optimally processed on the different (i. e., dedicated) machines $M = \{M_1, M_2, \ldots, M_{|M|}\}$. We next describe the scheduling problem $G|t_i = 1|C_{\max}$ along with our presentation of this problem by means of the mixed graph colouring $c(G)$.

In the problem $G|t_i = 1|C_{\max}$, a job $J_k \in J$ consists of a set $V^{(k)}$ of linearly ordered operations. The processing time t_i of each operation v_i in the set $V = \bigcup_{k=1}^{|J|} V^{(k)}$ is equal to 1; $t_i = 1$. Due to definition 1, we pre-

sent every job $J_k \in J$ as a union of path $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{r_k}}\right)$ in the directed subgraph $(V, A, \varnothing)$ and the chain $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{r_k}}\right)$ in subgraph $(V, \varnothing, E)$ of the mixed graph $G = (V, A, E)$, determining input data for the problem $G\left|t_i = 1\right|C_{\max}$. As a result, we define a vertex set $V = \bigcup_{k=1}^{|J|} V^{(k)}$ of the mixed graph $G = (V, A, E)$, a subset $E^* = \bigcup_{k=1}^{|J|} \left\{\left[v_{k_1}, v_{k_2}\right], \left[v_{k_2}, v_{k_3}\right], \ldots, \left[v_{k_{r_k-1}}, v_{k_{r_k}}\right]\right\}$ of the edge set $E \supseteq E^*$, and a subset A^* of the arc set A determined by the following implication:

$$\left[v_i, v_j\right] \in E^* \Rightarrow \left(v_i, v_j\right) \in A^*. \tag{1}$$

In the general shop-scheduling problem $G\left|t_i = 1\right|C_{\max}$, along with a linear order given on the set $V^{(k)}$ of all operations belonging to the same job $J_k \in J$, there are also given the precedence relations between operations belonging to different jobs in the set J. Let $A \backslash A^*$ denote a subset of set A such that implication (1) does not hold for each arc $\left(v_i, v_j\right) \in A \backslash A^*$. All the given precedence relations make up the precedence constraints.

In the problem $G\left|t_i = 1\right|C_{\max}$, a specified machine from the set $M = \left\{M_1, M_2, \ldots, M_{|M|}\right\}$ is required to process operation v_i from the set $V = \bigcup_{k=1}^{|J|} V^{(k)}$. Let $V_i = \left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\} \subseteq V$ denote a set of all operations processed on machine $M_i \in M$. Any pair of operations requiring the same machine $M_i \in M$ cannot be processed simultaneously [28; 29; 31–33]. We represent all such incompatibility constraints for processing operations $V_i \subseteq V$ on machine $M_i \in M$ (called capacity constraints) by cliques $\left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\}$ in the subgraph $\left(V, \varnothing, E \backslash E^*\right)$ of the mixed graph $G = (V, A, E)$ constructed for the problem $G\left|t_i = 1\right|C_{\max}$. The general shop-scheduling problem $G\left|t_i = 1\right|C_{\max}$ is to find a schedule for processing partially ordered operations $V = \bigcup_{i=1}^{|M|} V_i = \bigcup_{k=1}^{|J|} V^{(k)}$, whose length (makespan) $C_{\max} = \max\left\{C_1, C_2, \ldots, C_{|J|}\right\}$ is minimised among lengths of all feasible schedules. Hereafter, C_k denotes a completion time of the job $J_k \in J$. The minimisation of schedule length $C_{\max}$ for partially ordered operations V with unit processing times is reduced to the optimal colouring $c(G)$ of the mixed graph $G = (V, A, E)$, where the vertex set V is a set of operations, the arc set A determines the precedence constraints, and the edge set E determines the capacity constraints. More precisely, the union $A^* \bigcup E^*$ of the arc set A^* and the edge set E^* determines $|J|$ subsets $V^{(k)}$ of linearly ordered operations of the jobs $J_k \in J$. The subset $E \backslash E^*$ of edges determines $|M|$ cliques $\left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\}$ in the graph $\left(V, \varnothing, E \backslash E^*\right)$, where all operations $\left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\}$ are processed on machine $M_i \in M$. The precedence relations between operations belonging to different jobs are determined in the directed subgraph $\left(V, A \backslash A^*, \varnothing\right)$ of the mixed graph $G = (V, A, E)$.

To illustrate the above reduction of the problem $G\left|t_i = 1\right|C_{\max}$ to the optimal colouring $c(G)$, we consider example 1 of the problem $G\left|t_i = 1\right|C_{\max}$ with four jobs and six machines (fig. 1). Let the machine set $M = \left\{M_1, M_2, \ldots, M_6\right\}$ have to process the job set $J = \left\{J_1, J_2, J_3, J_4\right\}$. Job $J_1 \in J$ consists of the set $V^{(1)} = \left\{v_1, v_2, v_3\right\}$ of linearly ordered operations. Job $J_1 \in J$ is represented by a union of the path $\left(v_1, v_2, v_3\right)$ in the digraph $(V, A, \varnothing)$ and the chain $\left(v_1, v_2, v_3\right)$ in the graph $(V, \varnothing, E)$. Job $J_2 \in J$ consists of the set $V^{(2)} = \left\{v_4, v_5, v_6, v_7, v_8\right\}$ of linearly ordered operations. Job $J_2 \in J$ is represented by a union of the path $\left(v_4, v_5, v_6, v_7, v_8\right)$ in the digraph $(V, A, \varnothing)$ and the chain $\left(v_4, v_5, v_6, v_7, v_8\right)$ in the graph $(V, \varnothing, E)$. Job $J_3 \in J$ consists of the set $V^{(3)} = \left\{v_9, v_{10}, v_{11}, v_{12}\right\}$ of linearly ordered operations. Job $J_3 \in J$ is represented by a union of the path $\left(v_9, v_{10}, v_{11}, v_{12}\right)$ in the digraph $(V, A, \varnothing)$ and the chain $\left(v_9, v_{10}, v_{11}, v_{12}\right)$ in the graph $(V, \varnothing, E)$. Job $J_4 \in J$ consists of the set $V^{(4)} = \left\{v_{13}, v_{14}, v_{15}\right\}$ of linearly ordered operations. Job $J_4 \in J$ is represented by a union of the path $\left(v_{13}, v_{14}, v_{15}\right)$ in the digraph $(V, A, \varnothing)$ and the chain $\left(v_{13}, v_{14}, v_{15}\right)$ in the graph $(V, \varnothing, E)$.

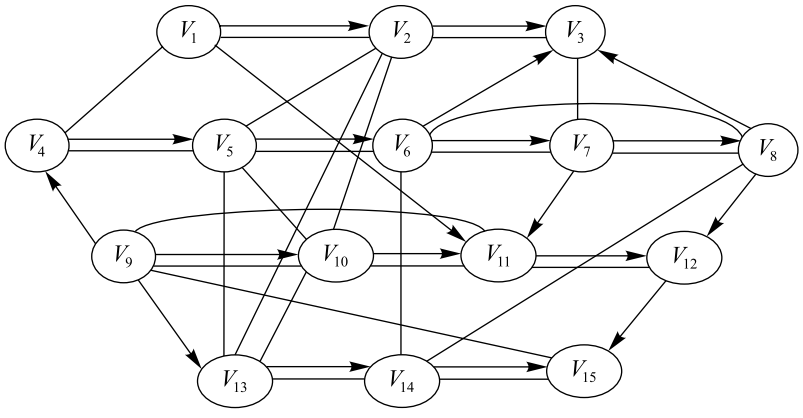

Fig. 1. Mixed graph $G=(V, A, E)$ determining example 1 of the problem $G|t_i=1|C_{\max}$ with four jobs and six machines, the optimal mixed graph colouring $c(G)$ being equivalent to example 1

Machine M_1 processes operations of the set $V_1=\{v_1, v_4\}$. The forbiddance to process operations from set V_1 simultaneously is represented by the clique $\{v_1, v_4\}$ in graph $(V, \varnothing, E)$. Machine M_2 processes operations $V_2=\{v_2, v_5, v_{10}, v_{13}\}$. The forbiddance to process each pair of operations from set V_2 simultaneously is represented by the clique $\{v_2, v_5, v_{10}, v_{13}\}$ in graph $(V, \varnothing, E)$. Machine M_3 processes operations $V_3=\{v_3, v_7\}$. The forbiddance to process operations from set V_2 simultaneously is represented by the clique $\{v_3, v_7\}$ in graph $(V, \varnothing, E)$. Machine M_4 processes operations $V_3=\{v_9, v_{11}, v_{15}\}$. The forbiddance to process each pair of operations from set V_2 simultaneously is represented by the clique $\{v_9, v_{11}, v_{15}\}$ in graph $(V, \varnothing, E)$. Machine M_5 processes operations $V_5=\{v_6, v_8, v_{14}\}$. The forbiddance to process each pair of operations from set V_2 simultaneously is represented by the clique $\{v_6, v_8, v_{14}\}$ in graph $(V, \varnothing, E)$. Machine M_6 processes only one operation: $V_6=\{v_{12}\}$.

Let the precedence relations between operations of the set V belonging to different jobs of the set J be given as follows: $v_1 \to v_{11}$; $v_6 \to v_3$; $v_8 \to v_3$; $v_7 \to v_{11}$; $v_8 \to v_{12}$; $v_9 \to v_4$; $v_9 \to v_{13}$; $v_{12} \to v_{15}$. These precedence relations determine the following set of arcs: $A \backslash A^* = \{(v_1, v_{11}), (v_6, v_3), (v_8, v_3), (v_7, v_{11}), (v_8, v_{12}), (v_9, v_4), (v_9, v_{13}), (v_{12}, v_{15})\}$ in the mixed graph $G=(V, A, E)$ such that implication (1) does not hold for each arc in the set $A \backslash A^*$.

Similarly to the mixed graph, representing input data of a shop-scheduling problem without operation preemptions [21; 28; 29; 32], input data for example 1 of the problem $G|t_i=1|C_{\max}$ is given by the mixed graph $G=(V, A, E)$ depicted in fig. 1, where a set of all operations is represented by the vertex set $V=\bigcup\limits_{i=1}^{|M|} V_i = \bigcup\limits_{k=1}^{|J|} V^{(k)}$. The precedence constraints and capacity constraints are represented by a union of arc set A and edge set E.

Based on the above reduction of the general shop-scheduling problem $G|t_i=1|C_{\max}$ to the colouring $c(G)$ of a suitable mixed graph $G=(V, A, E)$, one can derive the following correspondence of terms used in the optimal colouring $c(G)$ of the mixed graph $G=(V, A, E)$ and terms used in the general shop-scheduling problem $G|t_i=1|C_{\max}$:

$$\{\text{vertex } v_i \in V\} \Leftrightarrow \{\text{non-preemptive unit-time operation } v_i \in V\};$$

$$\{\text{vertices on path (on chain)} \left(v_{k_1}, v_{k_2}, \ldots, v_{k_{1_{|V^{(k)}|}}}\right) \text{ in digraph } (V, A^*, \varnothing) \text{ (in graph } (V, \varnothing, E^*))\} \Leftrightarrow$$

$$\Leftrightarrow \{\text{set } V^{(k)} = \left\{v_{k_1}, v_{k_2}, \ldots, v_{k_{|V^{(k)}|}}\right\} \text{ of linearly ordered operations of the job } J_k \in J\};$$

$$\{\text{precedence relations between operations belonging to different jobs}\} \Leftrightarrow \{\text{set of arcs } A \backslash A^* \text{ in digraph } (V, A \backslash A^*, \varnothing)\};$$

$$\{\text{clique } \left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\} \text{ in graph } \left(V, \varnothing, E \backslash E^*\right)\} \Leftrightarrow \{\text{operations } V_i = \left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\} \text{ processed on machine } M_i \in M\};$$

{a colouring $c(G)$ of the mixed graph $G=(V, A, E)$} $\Leftrightarrow$ {a feasible schedule for the problem $G|t_i=1|C_{\max}$};

{an optimal mixed graph colouring $c(G)$} $\Leftrightarrow$ {an optimal schedule for the problem $G|t_i=1|C_{\max}$};

{the chromatic number $\chi(G)$} $\Leftrightarrow$ {the optimal value of makespan $C_{\max}$}.

The above correspondence of terms used in the optimal colouring $c(G)$ of the mixed graph $G=(V, A, E)$ and those used in the general shop-scheduling problem $G|t_i=1|C_{\max}$ implies the following claim.

Lemma 1. *Any general shop-scheduling problem* $G|t_i=1|C_{\max}$ *may be represented as an optimal mixed graph colouring* $c(G)$ *of a suitable mixed graph* $G=(V, A, E)$.

However, it is easy to see that an inverse claim to lemma 1 is not correct.

An optimal schedule for example 1 is determined by the following optimal colouring $c(G)$ of the mixed graph $G=(V, A, E)$: $c(v_1)=2$, $c(v_2)=4$, $c(v_3)=5$, $c(v_4)=1$, $c(v_5)=2$, $c(v_6)=3$, $c(v_7)=4$, $c(v_8)=5$, $c(v_9)=1$, $c(v_{10})=3$, $c(v_{11})=4$, $c(v_{12})=5$, $c(v_{13})=1$, $c(v_{14})=4$, $c(v_{15})=5$. This colouring $c(G)$ of the mixed graph $G=(V, A, E)$ is optimal, i. e., $\chi(G)=5$. Indeed, the optimality of the schedule determined by the mixed graph colouring $c(G)$ follows from the fact that there is a job $J_2 \in J$ consisting of five operations $V^{(2)}=\{v_4, v_5, v_6, v_7, v_8\}$, which implies the following non-strict inequality: $\chi(G)\geq 5$.

Job-shop scheduling with unit-time operations to minimise makespan. Article [2] with definition 1 of the mixed graph colouring $c(G)$ was published in Russian in 1976 along with other articles published before 1997. In 1997, another mixed graph colouring (called a strict mixed graph colouring $c_<(G)$) has been introduced in article [19] published in English.

Definition 2 [19]. An integer-valued function $c_<: V \to \{1, 2, \ldots, t\}$ is a strict colouring of the mixed graph $G=(V, A, E)$, if inequality $c_<(v_i) < c_<(v_j)$ holds for each arc $(v_i, v_j) \in A$ and $c(v_p) \neq c(v_q)$ for each edge $[v_p, v_q] \in E$. A strict mixed graph colouring $c_<(G)$ is optimal, if it uses a minimal possible number $\chi_<(G)$ of different colours $c_<(v_i) \in \{1, 2, \ldots, t\}$. A minimal number $\chi_<(G)$ is a strict chromatic number of the mixed graph $G=(V, A, E)$.

It is clear that one can use a colouring $c(G)$ (definition 1) instead of a strict colouring $c_<(G)$ (definition 2) for every specific mixed graph $G=(V, A, E)$ such that the following implication (2) holds for each arc $(v_i, v_j) \in A$:

$$(v_i, v_j) \in A \Rightarrow [v_i, v_j] \in E. \quad (2)$$

Remark 1. A strict colouring $c_<(G)$ of the mixed graph $G=(V, A, E)$ is a special case of the colouring $c(G)$, if it is assumed that each inclusion $(v_i, v_j) \in A$ implies the inclusion $[v_i, v_j] \in E$ in the mixed graph $G=(V, A, E)$ to be coloured.

Due to remark 1, one can add edge $[v_i, v_j]$ to the mixed graph $G=(V, A, E)$ for each arc $(v_i, v_j) \in A$ such that implication (2) does not hold. Obviously, any strict colouring $c_<(G)$ of the mixed graph $G=(V, A, E)$ is a strict colouring $c_<(G^+)$ of the mixed graph $G^+=(V, A, E^+)$ constructed via adding all above edges $[v_i, v_j]$. Furthermore, strict mixed graph colourings $c_<(G)$ and $c_<(G^+)$ are the same as a mixed graph colouring $c(G^+)$.

The connection of the strict mixed graph colouring $c_<(G)$ and the job-shop scheduling problem $J|t_i=1|C_{\max}$ is studied in [4–8]. The job-shop scheduling problem $J|t_i=1|C_{\max}$ is a special case of the general shop-scheduling problem $G|t_i=1|C_{\max}$, if there are no precedence relations between operations belonging to different jobs (see [28–30]).

In article [4], it is shown that a mixed graph $G=(V, A, E)$ determining a job-shop scheduling problem $J|t_i=1|C_{\max}$ has the following mandatory properties.

Property 1. The partition $(V, \varnothing, E)=(V_1, \varnothing, E_1)\bigcup(V_2, \varnothing, E_2)\bigcup\ldots\bigcup(V_m, \varnothing, E_m)$ holds, where the subgraph $(V_k, \varnothing, E_k)$ of the mixed graph $G=(V, A, E)$ is a complete graph for each $k \in \{1, 2, \ldots, m\}$.

Property 2. The following partition $(V, A, \varnothing) = \left(V^{(1)}, A^{(1)}, \varnothing\right) \bigcup \left(V^{(2)}, A^{(2)}, \varnothing\right) \bigcup \ldots \bigcup \left(V^{(r)}, A^{(r)}, \varnothing\right)$ holds, where each directed subgraph $\left(V^{(k)}, A^{(k)}, \varnothing\right)$ of the mixed graph $G = (V, A, E)$ is a path $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{r_k}}\right)$ for $k \in \{1, 2, \ldots, r\}$.

Property 1 (property 2) means that the subgraph $(V, \varnothing, E)$ of the mixed graph $G = (V, A, E)$ is a union of disjoint complete graphs (the directed subgraph $(V, A, \varnothing)$ is a union of disjoint paths, respectively). In the job-shop scheduling problem $J|t_i = 1|C_{\max}$, numbers m and r denote the cardinality of the machine set $M = \{M_1, M_2, \ldots, M_{|M|}\}$, $m = |M|$, and the cardinality of job set $J = \{J_1, J_2, \ldots, J_{|J|}\}$, $r = |J|$. Property 2 implies that if the inclusion $v_i \in V^{(k)}$ holds, operation v_i belongs to the job $J_k \in J$, and vice versa (see definition 2). A job $J_k \in J$ consisting of a set $V^{(k)}$ of linearly ordered operations is represented as path $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{r_k}}\right)$ in digraph $(V, A, \varnothing)$. Operations $V^{(k)}$ have to be processed in the order determined by path $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{r_k}}\right)$. Property 1 means that if the inclusion $v_i \in V_k$ holds, operation v_i has to be processed on machine $M_k \in M$. Due to definition 2 and property 1, each machine $M_k \in M$ can process at most one operation within any unit-time interval from the following set:

$$\{[0, 1], (1, 2], (2, 3], \ldots, (t-1, t]\}. \tag{3}$$

An optimal strict colouring $c_< : V \to \{1, 2, \ldots, \chi_<(G)\}$ of the mixed graph $G = (V, A, E)$ determines an assignment of operations V to a minimal number of the following intervals:

$$\{[0, 1], (1, 2], (2, 3], \ldots, (\chi_<(G) - 1, \chi_<(G)]\}. \tag{4}$$

An assignment of operations V to the minimal number of unit-time intervals (4) is optimal since it determines a makespan optimal schedule for processing operations V, whose length is equal to the strict chromatic number $\chi_<(G)$ of the mixed graph $G = (V, A, E)$ determining an example of the unit-time minimum-length job-shop scheduling problem $J|t_i = 1|C_{\max}$. Properties 1 and 2 define the usual assumptions used in scheduling theory [28–30] in terms of graph theory [26; 27]. The following lemma 2 is proved in article [4].

Lemma 2 [4]. *Any individual job-shop scheduling problem* $J|t_i = 1|C_{\max}$ *is equivalent to an optimal strict colouring* $c_<(G)$ *of a suitable mixed graph* $G = (V, A, E)$ *possessing both properties 1 and 2, and vice versa.*

The proof of lemma 2 is based on the following correspondence of terms used in the strict mixed graph colouring $c_<(G)$ and those used in the job-shop problem $J|t_i = 1|C_{\max}$:

{vertex $v_i \in V$} $\Leftrightarrow$ {non-preemptive unit-time operation $v_i \in V$};

{vertices on path $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{|V^{(k)}|}}\right)$ in digraph $(V, A, \varnothing)$} $\Leftrightarrow$ {set $V^{(k)} = \left\{v_{k_1}, v_{k_2}, \ldots, v_{k_{|V^{(k)}|}}\right\}$ of linearly ordered operations of the job $J_k \in J$};

{clique $\left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\}$ in graph $(V, \varnothing, E)$} $\Leftrightarrow$ {operations $V_i = \left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\right\}$ processed on machine $M_i \in M$};

{a strict mixed graph colouring $c_<(G)$} $\Leftrightarrow$ {a schedule for the problem $G|t_i = 1|C_{\max}$};

{an optimal strict mixed graph colouring $c_<(G)$} $\Leftrightarrow$ {an optimal schedule for the problem $G|t_i = 1|C_{\max}$};

{the strict chromatic number $\chi_<(G)$} $\Leftrightarrow$ {the optimal value of makespan $C_{\max}$}.

To illustrate lemma 2, we consider example 2 of the problem $J|t_i = 1|C_{\max}$, which is the same as already considered example 1 of the problem $G|t_i = 1|C_{\max}$ with only one exception that there is no precedence constraint between operations belonging to different jobs in the set J. In other words, it is assumed that $A \backslash A^* = \varnothing$. It is clear that a strict colouring $c_<(G)$ of the mixed graph $G = (V, A, E)$ depicted in fig. 2 determines a schedule existing for example 2. Obviously, the mixed graph $G = (V, A, E)$ depicted in fig. 2 possesses both properties 1 and 2. This mixed graph $G = (V, A, E)$ is a subgraph of the mixed graph depicted in fig. 1.

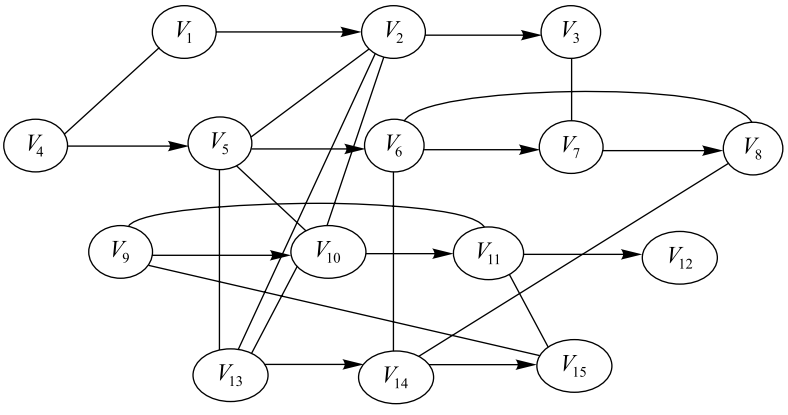

Fig. 2. Mixed graph $G=(V, A, E)$ determining example 2 of the problem $J|t_i=1|C_{\max}$ with four jobs and six machines, the optimal strict mixed graph colouring $c_<(G)$ being equivalent to example 2

An optimal schedule for example 2 is determined by the following strict mixed graph colouring $c_<(G)$: $c_<(v_1)=2$, $c_<(v_2)=4$, $c_<(v_4)=1$, $c_<(v_6)=3$, $c_<(v_7)=4$, $c_<(v_8)=5$, $c_<(v_9)=1$, $c_<(v_{10})=3$, $c_<(v_{11})=4$, $c_<(v_{12})=5$, $c_<(v_{13})=1$, $c_<(v_{14})=4$, $c_<(v_{15})=5$. This strict colouring $c_<(G)$ is optimal, i. e., $\chi_<(G)=5$, due to the existence of a job $J_2 \in J$ with five operations $V^{(2)}=\{v_4, v_5, v_6, v_7, v_8\}$ implying the following non-strict inequality: $\chi_<(G)\geq 5$.

It is important to highlight that there exists a general shop-scheduling problem $G|t_i=1|C_{\max}$ which cannot be represented as the optimal strict colouring $c_<(G)$ of a mixed graph $G=(V, A, E)$. This shortage of a strict mixed graph colouring to represent a general shop-scheduling problem occurs since the strict inequality $c_<(v_i)<c_<(v_j)$ must hold for each arc $(v_i, v_j)\in A$ in the colouring $c_<(G)$, and therefore, a strict mixed graph colouring cannot define a precedence relation $v_i \to v_j$ on the operations v_i and v_j belonging to different jobs in the set J.

Remark 2. There are general shop-scheduling problems $G|t_i=1|C_{\max}$, which cannot be represented as optimal strict colourings $c_<(G)$ of the suitable mixed graphs $G=(V, A, E)$.

In the following section, we introduce a new class of the scheduling problems that is more general than the classes of the problems $G|t_i=1|C_{\max}$ and $J|t_i=1|C_{\max}$ considered in this section. Based on the newly introduced class of the scheduling problems, we prove that an optimal colouring $c(G)$ of any colourable mixed graph $G=(V, A, E)$ is equivalent to an appropriate optimal unit-time minimum-length scheduling problem, and vice versa.

Unit-time scheduling partially ordered multi-processor tasks

Contrary to the scheduling problems studied in the previous section, where each operation has to be processed on a single machine, in the scheduling system with multi-processor tasks (MPT), a task may require either one processor (machine) or several processors during the complete period of processing the task [29; 33–36]. As usual, two tasks (operations) requiring at least one common processor (machine) cannot be processed simultaneously.

Chapter 10 of the book [29, p. 264–283] studies a general shop minimum-length scheduling problem $GMPT|t_i=1|C_{\max}$ along with other scheduling problems $MPT|\beta|\gamma$ with multi-processor tasks [33–36]. The symbol G in the field α of the three-field notation $GMPT|t_i=1|C_{\max}$ specifies a task system with arbitrary precedence constraints given on the set $V=\{v_1, v_2, \ldots, v_n\}$ of the multi-processor tasks. In the problem $GMPT|t_i=1|C_{\max}$, it is needed to construct an optimal schedule for processing partially ordered multi-processor tasks $V=\{v_1, v_2, \ldots, v_n\}$ on the dedicated processors $M=\{M_1, M_2, \ldots, M_{|M|}\}$. The general shop-scheduling problem $G|t_i=1|C_{\max}$ is a special case of the problem $GMPT|t_i=1|C_{\max}$ since the processing of task $v_i \in V$ requires a single processor for the problem $G|t_i=1|C_{\max}$. In the general shop-scheduling problem $GMPT|t_i=1|C_{\max}$, a task $v_i \in V$ may be regarded as a job J_i including either one operation (task) $v_i \in V$ or more than one operation (several tasks from the set V). Let a simple job mean a job consisting only of one operation (task).

For any example of the problem $GMPT|t_i=1|C_{\max}$, one can construct a mixed graph $G=(V, A, E)$ such that an optimal colouring $c(G)$ of the mixed graph $G=(V, A, E)$ is equivalent to finding an optimal schedule for the problem $GMPT|t_i=1|C_{\max}$. The construction of such a mixed graph $G=(V, A, E)$ is analogous to the construction of the mixed graph $G=(V, A, E)$ determining input data for the problem $G|t_i=1|C_{\max}$ (see the previous section).

We next introduce a new class of the general shop-scheduling problems $GcMPT|t_i=1|C_{\max}$, which includes the problem $GMPT|t_i=1|C_{\max}$ as a special case studied in chapter 10 of the book [29]. More precisely, in the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$, it is required that a subset $V(k)=\left\{v_{k_1}, v_{k_2}, \ldots, v_{k_{|V(k)|}}\right\}$ of the tasks $V=\{v_1, v_2, \ldots, v_n\} \supseteq V(k)$ must be processed simultaneously in any feasible schedule. It is easy to see that the latter requirement may be represented by a circuit $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{|V(k)|}}, v_{k_1}\right)$ in the directed subgraph $(V, A_c, \varnothing)$ of the mixed graph $G=(V, A, E)$, which presents input data of the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$, where the set A of the arcs includes the following subset: $A'_c=\left\{\left(v_{k_1}, v_{k_2}\right), \left(v_{k_2}, v_{k_3}\right), \ldots, \left(v_{k_{|V(k)|-1}}, v_{k_{|V(k)|}}\right), \left(v_{k_{|V(k)|}}, v_{k_1}\right)\right\} \subseteq A$.

Let the input data for the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ include w subsets $V(1), V(2), \ldots, V(w)$ of the tasks such that every subset $V(k)=\left\{v_{k_1}, v_{k_2}, \ldots, v_{k_{|V(k)|}}\right\}$ of the tasks $V=\{v_1, v_2, \ldots, v_n\}$ must be processed simultaneously in any feasible schedule, where $k \in \{1, 2, \ldots, w\}$. Then, we determine the following subset of arcs:

$$A_c=\bigcup_{k=1}^{w}\left\{\left(v_{k_1}, v_{k_2}\right), \left(v_{k_2}, v_{k_3}\right), \ldots, \left(v_{k_{|V(k)|-1}}, v_{k_{|V(k)|}}\right), \left(v_{k_{|V(k)|}}, v_{k_1}\right)\right\}. \quad (5)$$

Similarly as in the previous section, one can establish the correspondence of terms used in the optimal colouring $c(G)$ of the mixed graph $G=(V, A, E)$ with $A_c \subseteq A$ and those used in the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ see the table.

Obviously, every instance of the problem $GcMPT|t_i=1|C_{\max}$ uniquely defines a mixed graph $G=(V, A, E)$ determining input data for this instance. Therefore, to describe an instance of the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$, it is sufficient to define a mixed graph $G=(V, A, E)$, which determines input data for this instance of the scheduling problem. In what follows, such an instance of the problem $GcMPT|t_i=1|C_{\max}$ will be called the problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$.

The correspondence of terms used in the mixed graph colouring $c(G)$ and those used in the problem $GcMPT|t_i=1|C_{\max}$ on mixed graph $G=(V, A, E)$

Terms of the mixed graph colouring $c(G)$	Terms of the problem $GcMPT\|t_i=1\|C_{\max}$
Vertex $v_i \in V$	Unit-time task $v_i \in V$ (unit-time operation of the job)
Vertices on the path (on the chain, respectively) $\left(v_{k_1}, v_{k_2}, \ldots, v_{k_{1_{\lvert V(k)\rvert}}}\right)$ in the digraph $(V, A^*, \varnothing)$ (in the graph $(V, \varnothing, E^*)$)	Set $V^{(k)}=\left\{v_{k_1}, v_{k_2}, \ldots, v_{k_{\lvert V(k)\rvert}}\right\}$ of linearly ordered operations (tasks) of the job $J_k \in J$
Clique $\left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{\lvert V_i\rvert}}\right\}$ in the graph $(V, \varnothing, E \backslash E^*)$	All tasks $V_i=\left\{v_{i_1}, v_{i_2}, \ldots, v_{i_{\lvert V_i\rvert}}\right\}$ processed on the same machine (processor) $M_i \in M$
Set of arcs $A \backslash A^*$ in the digraph $(V, A \backslash A^*, \varnothing)$	Precedence relations given between tasks (operations) belonging to different jobs of the set J

Ending table

Terms of the mixed graph colouring $c(G)$	Terms of the problem $GcMPT\left\|t_i=1\right\|C_{\max}$
Set of arcs $A\backslash A^*$ in the digraph $\left(V,\ A\backslash A^*,\ \varnothing\right)$	Precedence constraints given on the set of tasks V
Circuit $\left(v_{k_1},\ v_{k_2},\ \ldots,\ v_{k_{\left\|V(k)\right\|}},\ v_{k_1}\right)$ in the digraph $(V,\ A,\ \varnothing)$, where $A_c \subseteq A$	Tasks $V(k)=\left\{v_{k_1},\ v_{k_2},\ \ldots,\ v_{k_{\left\|V(k)\right\|}}\right\}\subseteq V$ that must be processed simultaneously
A mixed graph colouring $c(G)$ of the mixed graph $G=(V,\ A,\ E)$	A feasible schedule for the problem $GcMPT\left\|t_i=1\right\|C_{\max}$
An optimal mixed graph colouring $c(G)$ of the mixed graph $G=(V,\ A,\ E)$	An optimal schedule for the problem $GcMPT\left\|t_i=1\right\|C_{\max}$
The chromatic number $\chi(G)$	The optimal value of makespan $C_{\max}$

Due to the correspondence of terms used in the colouring $c(G)$ and those used in the equivalent problem $GcMPT\left|t_i=1\right|C_{\max}$ on the mixed graph $G=(V,\ A,\ E)$, one can derive lemma 3.

Lemma 3. *Every general shop-scheduling problem* $GcMPT\left|t_i=1\right|C_{\max}$ *on the mixed graph* $G=(V,\ A,\ E)$ *is equivalent to an optimal mixed graph colouring* $c(G)$.

Contrary to the job-shop scheduling problem $J\left|t_i=1\right|C_{\max}$ having a feasible schedule for any input data, there are instances of the general shop-scheduling problem $G\left|t_i=1\right|C_{\max}$, which have no feasible schedules. To construct an instance of such an unsolvable individual general shop-scheduling problem $G\left|t_i=1\right|C_{\max}$, we add the precedence relation $v_5 \to v_9$ to the input data of example 1 depicted in fig. 1. We call this modified example as example 1^* and show that there is no feasible schedule for example 1^* due to the existence of the circuit (v_4, v_5, v_9, v_4) in the digraph $(V,\ A,\ \varnothing)$ and the edge $\left[v_4, v_5\right]$ in the graph $\left(V,\ \varnothing,\ E^*\right)$. On the one hand, all tasks in the set $\left\{v_4, v_5, v_9\right\}$ must be processed simultaneously due to the circuit (v_4, v_5, v_9, v_4). On the other hand, two tasks v_4 and v_5 cannot be processed simultaneously due to the edge $\left[v_4, v_5\right] \in E^* \subset E$. This contradiction implies that there is no feasible schedule for example 1^*. Since the general shop-scheduling problem $G\left|t_i=1\right|C_{\max}$ is a special case of the general shop-scheduling problem $GcMPT\left|t_i=1\right|C_{\max}$, there are similar instances of the problem $GcMPT\left|t_i=1\right|C_{\max}$ such that no feasible schedules exist.

We prove the following criterion for the existence of a feasible schedule for the general shop-scheduling problem $GcMPT\left|t_i=1\right|C_{\max}$ on the mixed graph $G=(V,\ A,\ E)$.

Theorem 2. *A feasible schedule for the general shop-scheduling problem* $GcMPT\left|t_i=1\right|C_{\max}$ *on the mixed graph* $G=(V,\ A,\ E)$ *exists, if and only if the digraph* $(V,\ A,\ \varnothing)$ *has no circuit containing adjacent vertices in the graph* $(V,\ \varnothing,\ E)$.

Proof. Due to lemma 3, a general shop-scheduling problem $GcMPT\left|t_i=1\right|C_{\max}$ on the mixed graph $G=(V,\ A,\ E)$ is equivalent to optimal colouring $c(G)$ of the mixed graph $G=(V,\ A,\ E)$. A mixed graph $G=(V,\ A,\ E)$ with $A \neq \varnothing$ and $E \neq \varnothing$ may be uncolourable, i. e., there is no colouring $c(G)$ for the mixed graph $G=(V,\ A,\ E)$. Furthermore, theorem 1 establishes a criterion for the existence of a colouring $c(G)$ for the mixed graph $G=(V,\ A,\ E)$ and this criterion directly proves theorem 2.

To illustrate lemma 3 and theorem 2, we consider two examples of the general shop-scheduling problem $GcMPT\left|t_i=1\right|C_{\max}$ with two non-simple jobs J_1 and J_2, eleven multi-processor tasks $V=\left\{v_1, v_2, \ldots, v_{11}\right\}$, seven machines $M=\left\{M_1, M_2, \ldots, M_7\right\}$, and three tasks $\left\{v_1, v_4, v_8\right\}$, which must be processed simultaneously in any feasible schedule. The mixed graph $G=(V,\ A,\ E)$ depicted in fig. 3 determines input data for example 3.

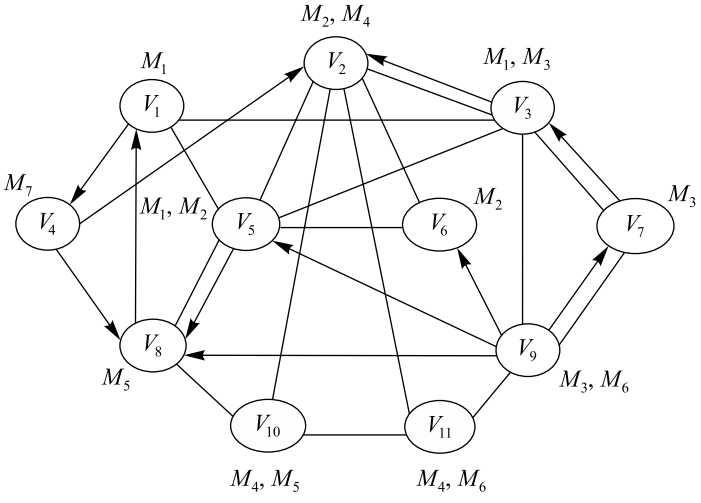

Fig. 3. Mixed graph $G=(V,\ A,\ E)$ determining the problem $GcMPT\left|t_i=1\right|C_{\max}$ with eleven tasks and seven machines, the optimal mixed graph colouring $c(G)$ being equivalent to example 3

In example 3, machine M_1 has to process three tasks of the set $V_1=\{v_1,\ v_3,\ v_5\}$. The forbiddance to process any pair of tasks from the set V_1 simultaneously is represented by the clique $\{v_1,\ v_3,\ v_5\}$ in the graph $(V,\ \varnothing,\ E\backslash E^*)$. Machine M_2 has to process three tasks of the set $V_2=\{v_2,\ v_5,\ v_6\}$. The forbiddance to process any pair of tasks from the set V_2 simultaneously is represented by the clique $\{v_2,\ v_5,\ v_6\}$ in the graph $(V,\ \varnothing,\ E\backslash E^*)$. Machine M_3 has to process three tasks from the set $V_3=\{v_3,\ v_7,\ v_9\}$. The forbiddance to process any pair of tasks from the set V_3 simultaneously is represented by the clique $\{v_3,\ v_7,\ v_9\}$ in the graph $(V,\ \varnothing,\ E\backslash E^*)$. Machine M_4 has to process three tasks of the set $V_4=\{v_2,\ v_{10},\ v_{14}\}$. The forbiddance to process any pair of tasks from the set V_4 simultaneously is represented by the clique $\{v_2,\ v_{10},\ v_{14}\}$ in the graph $(V,\ \varnothing,\ E\backslash E^*)$. Machine M_5 has to process two tasks of the set $V_5=\{v_8,\ v_{10}\}$. The forbiddance to process tasks from the set V_5 simultaneously is represented by the clique $\{v_8,\ v_{10}\}$ in the graph $(V,\ \varnothing,\ E\backslash E^*)$. Machine M_6 has to process two tasks of the set $V_6=\{v_9,\ v_{11}\}$. The forbiddance to process tasks from the set V_6 simultaneously is represented by the clique $\{v_9,\ v_{11}\}$ in the graph $(V,\ \varnothing,\ E\backslash E^*)$. Machine M_7 has to process one task: $V_7=\{v_4\}$. All machines, which are used for processing the task $v_i\in V$, are presented near vertex v_i in fig. 3.

There are two jobs J_1 and J_2, which are not simple. Job $J_1\in J$ consists of the set $V^{(1)}=\{v_9,\ v_7,\ v_3,\ v_2\}$ of linearly ordered tasks (operations). Thus, job $J_1\in J$ is represented by a union of the path $(v_9,\ v_7,\ v_3,\ v_2)$ in the digraph $(V,\ A^*,\ \varnothing)$ and the chain $(v_9,\ v_7,\ v_3,\ v_2)$ in the graph $(V,\ \varnothing,\ E^*)$. Job $J_2\in J$ consists of the set $V^{(2)}=\{v_5,\ v_8\}$ of linearly ordered tasks (operations). Thus, job $J_1\in J$ is represented by a union of the path $(v_5,\ v_8)$ in the digraph $(V,\ A^*,\ \varnothing)$ and the chain $(v_5,\ v_8)$ in the graph $(V,\ \varnothing,\ E^*)$. Other precedence relations between tasks and operations belonging to different non-simple jobs are determined as follows: $v_1\to v_4$; $v_4\to v_2$; $v_4\to v_8$; $v_8\to v_1$; $v_9\to v_5$; $v_9\to v_6$; $v_9\to v_8$. All the precedence constraints determine the subset A of the arcs in the digraph $(V,\ A\backslash A^*,\ \varnothing)$, where $A\backslash A^*=\{(v_1,\ v_4),\ (v_4,\ v_2),\ (v_5,\ v_3),\ (v_4,\ v_8),\ (v_8,\ v_1),\ (v_9,\ v_5),\ (v_9,\ v_4),\ (v_9,\ v_6),\ (v_9,\ v_8)\}$.

In example 3, every job J_i from the set $J\backslash\{J_1,\ J_2\}$ is simple, i. e., job J_i consists of a single task that is identified with job J_i. In example 3, the set of tasks $V^{(1)}=\{v_1,\ v_4,\ v_8\}\subset V$ must be processed simultaneously in any feasible schedule. This requirement is determined by the circuit $(v_1,\ v_4,\ v_8,\ v_1)$ in the digraph $(V,\ A_c,\ \varnothing)$. Based on the correspondence of the terms (see the table), we construct the mixed graph $G=(V,\ A,\ E)$ depicted in fig. 3,

which determines the input data for example 3 of the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ with eleven multi-processor tasks, seven machines, two non-simple jobs, and three tasks, which must be processed simultaneously in any feasible schedule.

Due to theorem 2, there exists a feasible schedule for example 3 of the problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$. Due to lemma 3, an optimal schedule for example 3 may be determined by the following optimal colouring $c(G)$ of the mixed graph $G=(V, A, E)$ depicted in fig. 3: $c(v_1)=2$, $c(v_2)=4$, $c(v_3)=3$, $c(v_4)=2$, $c(v_5)=1$, $c(v_6)=2$, $c(v_7)=2$, $c(v_8)=2$, $c(v_9)=1$, $c(v_{10})=1$, $c(v_{11})=2$. This mixed graph colouring $c(G)$ is optimal due to lemma 3 since $\chi(G)=4$. Indeed, the optimality of the schedule determined by the mixed graph colouring $c(G)$ follows from the fact that there exists a job $J_1 \in J$ consisting of four unit-time operations $V^{(1)}=\{v_9, v_7, v_3, v_2\}$ that imply the non-strict inequality $\chi(G)\geq 4$.

Due to theorem 2, there are examples of the general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ such that no feasible schedules exist for them. We construct an example of such an unsolvable general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ as follows. We replace the precedence relation $v_9 \to v_8$ by the opposite relation $v_8 \to v_9$ in the input data of example 3 depicted in fig. 3. There is no feasible schedule for such a modified example 3 (we call it example 3^*), where the precedence relation $v_9 \to v_8$ is replaced by the relation $v_8 \to v_9$. On the one hand, all tasks from the set $\{v_5, v_8, v_9\}$ must be processed simultaneously due to the circuit (v_5, v_8, v_9, v_5). On the other hand, two tasks v_5 and v_8 cannot be processed simultaneously due to the edge $[v_5, v_8] \in E^* \subset E$. The contradiction obtained along with theorem 2 implies that there is no feasible schedule for example 3^* of the problem $GcMPT|t_i=1|C_{\max}$.

We next prove the following lemma, which is the inverse of lemma 3.

Lemma 4. *For any colourable mixed graph $G=(V, A, E)$, there exists a general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$, which is equivalent to finding an optimal colouring $c(G)$.*

Proof. We detect a set Ω of all circuits existing in the directed subgraph $(V, A, \varnothing)$ of the mixed graph $G=(V, A, E)$ and consider two possible cases: either $\Omega=\varnothing$ or $\Omega\neq\varnothing$.

Case 1. Let the set Ω be empty; $\Omega=\varnothing$. Then, one can construct the desired problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$ using the following algorithm.

Algorithm

Input: a mixed graph $G=(V, A, E)$ such that no circuit exists in the digraph $(V, A, \varnothing)$.

Output: a general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$, which is equivalent to finding an optimal colouring $c(G)$.

Step 1: partition the graph $(V, \varnothing, E)$ into (maximal) components as follows:

$$(V, \varnothing, E)=(V_1, \varnothing, E_1)\bigcup\ldots\bigcup(V_m, \varnothing, E_m)\bigcup(V_{m+1}, \varnothing, \varnothing)\bigcup\ldots\bigcup(V_{m+r}, \varnothing, \varnothing),$$

where the subgraph $(V_k, \varnothing, E_k)$ is a (maximal) component of the graph $(V, \varnothing, E)$ for each $k\in\{1, \ldots, m\}$ such that $|V_k|\geq 2$. The subgraph $(V_j, \varnothing, \varnothing)$ determines an isolated vertex for each index $j\in\{m+1, \ldots, m+r\}$. Denote this isolated vertex as follows: $\{v_{j_1}\}:=V_j$. Set $M=\varnothing$, $k=1$, $i=0$, $l_0=0$.

Step 2: IF $k=m+1$ **THEN GOTO** step 5 **ELSE** find all maximal (relative to inclusion) complete vertex-induced subgraphs $(V_k^1, \varnothing, E_k^1), \ldots, (V_k^{l_k}, \varnothing, E_k^{l_k})$ of the connected graph $(V_k, \varnothing, E_k)$. Set $r=1$, $i:=i+l_{k-1}+1$.

Step 3: FOR index i, supplement machine M_i with the already constructed machine set, i. e., $M:=M\cup\{M_i\}$. Establish that all tasks in the clique V_k^r of the connected graph $(V_k, \varnothing, E_k)$ must be processed on machine M_i, i. e., $V_k^r=V_i=\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\}$, where all tasks $\{v_{i_1}, v_{i_2}, \ldots, v_{i_{|V_i|}}\}$ must be processed on machine M_i in any feasible schedule. Set $i:=i+1$.

Step 4: IF $i=\sum_{h=0}^{k} l_h$ **THEN** set $k:=k+1$ **GOTO** step 2 **ELSE** set $r:=r+1$ **GOTO** step 3.

Step 5: FOR each index $j\in\{m+1, \ldots, m+r\}$, supplement machine M_{i+j-m} with the already constructed machine set M. Establish that task v_{j_1} with $V_j=\{v_{j_1}\}$, which is isolated in the graph $(V, \varnothing, E)$, must be processed on machine M_{i+j-m}. Establish that machine M_{i+j} must process only task v_{j_1}. Set $M:=M\cup\{M_{i+1}, \ldots, M_{i+r}\}$.

Step 6: FOR each arc (v_p, v_q) existing in the directed subgraph $(V, A, \varnothing)$ of the mixed graph $G=(V, A, E)$, introduce the precedence relation $v_p \to v_q$, which means that processing the task v_p must be completed before starting the task v_p in any feasible schedule.

Step 7: a general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ is constructed on the mixed graph $G=(V, A, E)$, where the precedence relations on the task set V are determined at step 6 and the machine set M is determined at step 3 and step 5 **STOP**.

Case 2. Let the set Ω be not empty, $\Omega = \varnothing$. Since the mixed graph $G=(V, A, E)$ is colourable, every circuit $(v_{k_1}, v_{k_2}, \ldots, v_{k_{|V(k)|}}, v_{k_1})$ in set Ω has no adjacent vertices in the subgraph $(V, \varnothing, E)$ of the mixed graph G (theorem 1). Therefore, all tasks $\{v_{k_1}, v_{k_2}, \ldots, v_{k_{|V(k)|}}\} =: V(k)$ must be processed simultaneously in any feasible schedule for the desired general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$, where the circuit $(v_{k_1}, v_{k_2}, \ldots, v_{k_{|V(k)|}}, v_{k_1})$ exists in the directed subgraph $(V, A, \varnothing)$.

Let $\Omega = \bigcup_{k=1}^{w} V(k) = \bigcup_{k=1}^{w} \left\{\left(v_{k_1}, v_{k_2}, v_{k_3}, \ldots, v_{k_{|V(k)|-1}}, v_{k_{|V(k)|}}, v_{k_1}\right)\right\}$. Then, we delete all arcs A_c defined in (5) from the mixed graph $G=(V, A, E)$ and apply the above algorithm to the obtained circuit-free mixed graph $G^0=(V, A\backslash A_c, E)$. As a result, the problem $GcMPT|t_i=1|C_{\max}$ is constructed on the mixed graph $G^0=(V, A\backslash A_c, E)$, which is equivalent to finding an optimal colouring $c(G^0)$. Consequently, the problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$ is equivalent to finding an optimal colouring $c(G)$. Lemma 4 is thus proved.

We apply the above constructive proof of lemma 2 to the mixed graph $G=(V, A, E)$ depicted in fig. 3 and obtain example 4 of the problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$ depicted in fig. 4. Note that examples 3 and 4 are different, e. g., all jobs are simple in example 4, while there are two non-simple jobs in example 3.

In general, it is easy to show that the proof of lemma 4 implies that for any colourable mixed graph $G=(V, A, E)$, one can construct a general shop-scheduling problem $GcMPT|t_i=1|C_{\max}$ on the mixed graph $G=(V, A, E)$, which is equivalent to finding an optimal colouring $c(G)$ and all jobs in the set J are simple.

Note that it is required to use non-simple jobs for some objective functions $\gamma = f(C_1, C_2, \ldots, C_{|J|})$, since only completion times C_i of the jobs $J_i \in J$ are their arguments. Hence, the completion times of some tasks may be ignored in the values of such objective functions. In particular, the total completion time $\gamma = \Sigma C_i = \sum_{i=1}^{|J|} C_i$ is such an objective function.

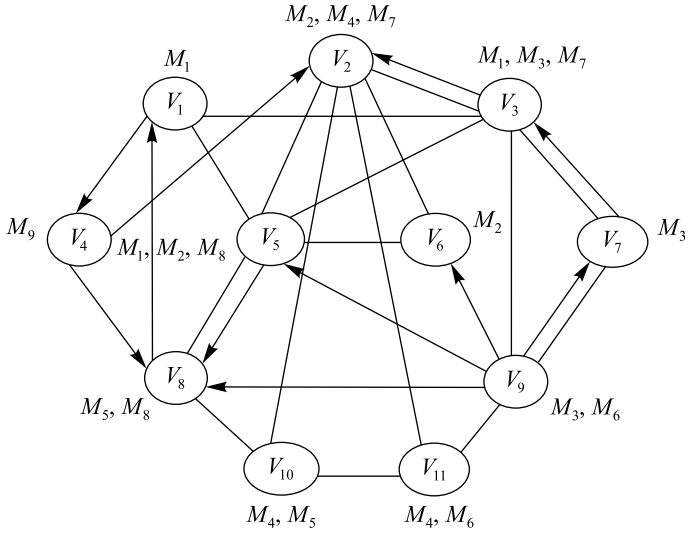

Fig. 4. Mixed graph $G=(V, A, E)$ determining the problem $GcMPT|t_i=1|C_{\max}$ with eleven tasks and nine machines, the optimal mixed graph colouring $c(G)$ being equivalent to example 4

Let $T = \{\tau_1, \tau_2, \ldots, \tau_{|T|}\}$ denote a set of all tasks (operations) in the scheduling problem $GcMPT|t_i = 1|C_{\max}$, and let $C(\tau_i)$ denote a completion time of the task $\tau_i \in T$. Due to the equalities $\max\{C_1, C_2, \ldots, C_{|J|}\} = C_{\max} = \max\{C(\tau_1), C(\tau_2), \ldots, C(\tau_{|T|})\}$, the completion time of any task cannot be ignored in the value of the objective function $\gamma = C_{\max}$. Therefore, only simple jobs may be used in the shop-scheduling problem with minimising makespan $C_{\max}$, where any non-simply job may be represented by the precedence relations on the task set T and the completion time $C(\tau_i)$ of each task $\tau_i \in T$ cannot be ignored in the value of $C_{\max}$.

Obviously, the following theorem combines lemmas 3 and 4.

Theorem 3. *Every general shop-scheduling problem* $GcMPT|t_i = 1|C_{\max}$ *on the mixed graph* $G = (V, A, E)$ *is equivalent to finding an optimal colouring* $c(G)$ *of the mixed graph* $G = (V, A, E)$. *Further, for any colourable mixed graph* $G = (V, A, E)$, *there exists an individual general shop-scheduling problem* $GcMPT|t_i = 1|C_{\max}$ *on the mixed graph* $G = (V, A, E)$, *which is equivalent to finding an optimal colouring* $c(G)$ *of the mixed graph* $G = (V, A, E)$.

To restrict a set of feasible schedules for a shop-scheduling problem $\alpha|\beta|\gamma$, which must be tested in order to minimise a value of the regular objective function γ [15], a finite set of semi-active schedules may be considered, since there exists an optimal semi-active schedule for a shop-scheduling problem $\alpha|\beta|\gamma$ with any regular objective function γ [28].

Definition 3 [28; 29]. A schedule is called semi-active, if no task (operation) can be processed earlier without violating a given constraint or changing the task (operation) processing order in the obtained schedule.

We remark that any colouring $c(G)$ of the mixed graph $G = (V, A, E)$ uniquely determines a strict order on the colours $c(v_j)$ of all vertices in the set V. Due to this remark, one can define a minimal colouring $c(G)$ of the mixed graph $G = (V, A, E)$ as follows.

Definition 4. A colouring $c(G)$ of the mixed graph $G = (V, A, E)$ is called minimal, if no colour $c(v_i)$ can be decreased without changing the order of colours $c(v_j)$ of the vertices in the set $V \setminus \{v_i\}$ in the obtained colouring $c'(G)$ of the mixed graph $G = (V, A, E)$.

Obviously, each semi-active schedule existing for the general shop-scheduling problem $GcMPT|t_i = 1|C_{\max}$ on the mixed graph $G = (V, A, E)$ uniquely determines a minimal colouring $c(G)$ of the mixed graph $G = (V, A, E)$, and vice versa. Hence, we obtain theorem 4.

Theorem 4. *There exists a one-to-one correspondence between all minimal colourings* $c(G)$ *of the mixed graph* $G = (V, A, E)$ *and all semi-active schedules existing for the general shop-scheduling problem* $GcMPT|t_i = 1|C_{\max}$ *on the mixed graph* $G = (V, A, E)$.

We used both graph terminology and scheduling one for the above problems. However, it is possible to describe most presented results either using only graph terminology or using only scheduling terminology.

Conclusion

We introduced a new class of general shop-scheduling problems $GcMPT|t_i = 1|C_{\max}$ for finding optimal schedules for partially ordered multi-processor tasks with unit processing times. Contrary to a classical shop-scheduling problem, several machines are required to process a task in the problem $GcMPT|t_i = 1|C_{\max}$. It is also required that a subset of tasks must be processed simultaneously in any feasible schedule. We proved theorem 3 showing that an optimal colouring $c(G)$ of any mixed graph $G = (V, A, E)$ is equivalent to the general shop-scheduling problem $GcMPT|t_i = 1|C_{\max}$, and vice versa. Hence, many terms used in scheduling theory (such as schedule, job, machine, processor, operation, task, processing time and makespan) may be considered as usual terms used in mixed graph colourings $c(G)$.

Due to theorems 3 and 4, most results that have been proven so far [4; 5; 7; 8; 14; 28; 29; 33–36] and will be proven in future for the scheduling problem $GcMPT|p_i = 1|C_{\max}$ and for its special cases have analogous results for optimal colourings $c(G)$ of the appropriate mixed graphs $G = (V, A, E)$. Conversely, most results that have been proven so far [2; 4; 5; 7; 8; 19; 25–27] and will be proven in future for optimal mixed graph colourings $c(G)$ have analogous results for scheduling problems $GcMPT|t_i = 1|C_{\max}$.

References

1. Wan L, Mei J, Du J. Two-agent scheduling of unit processing time jobs to minimize total weighted completion time and total weighted number of tardy jobs. *European Journal of Operational Research.* 2021;290(1):26–35. DOI: 10.1016/j.ejor.2020.07.064.

2. Sotskov YN, Tanaev VS. [A chromatic polynomial of a mixed graph]. *Izvestiya Akademii nauk BSSR. Seriya fiziko-matematicheskikh nauk.* 1976;6:20–23. Russian.

3. Karp RM. Reducibility among combinatorial problems. In: Miller RE, Thatcher JW, Bohlinger JD, editors. *Complexity of computer computations*. Boston: Springer; 1972. p. 85–103. DOI: 10.1007/978-1-4684-2001-2_9.

4. Sotskov YN, Dolgui A, Werner F. Mixed graph coloring for unit-time job-shop scheduling. *International Journal of Mathematical Algorithms.* 2001;2:289–323.

5. Sotskov YN, Tanaev VS, Werner F. Scheduling problems and mixed graph colorings. *Optimization.* 2002;51(3):597–624. DOI: 10.1080/0233193021000004994.

6. Al-Anzi FS, Sotskov YN, Allahverdi A, Andreev GV. Using mixed graph coloring to minimize total completion time in job shop scheduling. *Applied Mathematics and Computation.* 2006;182(2):1137–1148. DOI: 10.1016/j.amc.2006.04.063.

7. Kouider A, Ait Haddadène H, Ourari S, Oulamara A. Mixed graph colouring for unit-time scheduling. *International Journal of Production Research*. 2017;55(6):1720–1729. DOI: 10.1080/00207543.2016.1224950.

8. Kouider A, Ait Haddadène H, Oulamara A. On minimization of memory usage in branch-and-bound algorithm for the mixed graph coloring: application to the unit-time job shop scheduling. *Computer & Operations Research*. 2019;4967:1001–1008.

9. Lenstra JK, Rinnooy Kan AHG. Computational complexity of discrete optimization problems. *Annals of Discrete Mathematics.* 1979;4:121–140. DOI: 10.1016/S0167-5060(08)70821-5.

10. Sotskov YN. Complexity of optimal scheduling problems with three jobs. *Cybernetics.* 1990;26(5):686–692. DOI: 10.1007/BF01068549.

11. Sotskov YN. The complexity of shop-scheduling problems with two or three jobs. *European Journal of Operational Research*. 1991;53(3):326–336. DOI: 10.1016/0377-2217(91)90066-5.

12. Sotskov YN, Shakhlevich NV. NP-hardness of shop-scheduling problems with three jobs. *Discrete Applied Mathematics*. 1995;59(3):237–266. DOI: 10.1016/0166-218X(95)80004-N.

13. Kravchenko SA, Sotskov YN. Optimal makespan schedule for three jobs on two machines. *Mathematical Methods of Operations Research*. 1996;43(2):233–238. DOI: 10.1007/BF01680374.

14. Gonzalez T. Unit execution time shop problems. *Mathematics of Operations Research*. 1982;7(1):57–66. DOI: 10.1287/moor.7.1.57.

15. Brucker P, Kravchenko SA, Sotskov YN. On the complexity of two machine job-shop scheduling with regular objective functions. *OR Spektrum*. 1997;19:5–10. DOI: 10.1007/BF01539799.

16. Shakhlevich NV, Sotskov YN. Scheduling two jobs with fixed and nonfixed routes. *Computing*. 1994;52(1):17–30. DOI: 10.1007/BF02243393.

17. Shakhlevich NV, Sotskov YN, Werner F. Shop-scheduling problems with fixed and non-fixed machine orders of the jobs. *Annals of Operations Research*. 1999;92:281–304. DOI: 10.1023/A:1018943016617.

18. Damaschke P. Parameterized mixed graph coloring. *Journal of Combinatorial Optimization.* 2019;38:326–374. DOI: 10.1007/s10878-019-00388-z.

19. Hansen P, Kuplinsky J, de Werra D. Mixed graph colorings. *Mathematical Methods of Operations Research*. 1997;45:145–160. DOI: 10.1007/BF01194253.

20. Kruger K, Sotskov YN, Werner F. Heuristics for generalized shop scheduling problems based on decomposition. *International Journal of Production Research*. 1998;36(11):3013–3033. DOI: 10.1080/002075498192265.

21. Sotskov YN. Software for production scheduling based on the mixed (multi)graph approach. *Computing & Control Engineering Journal.* 1996;7(5):240–246. DOI: 10.1049/CCE:19960509.

22. Sotskov YN. Mixed multigraph approach to scheduling jobs on machines of different types. *Optimization*. 1997;42(3):245–280. DOI: 10.1080/02331939708844361.

23. de Werra D. Restricted coloring models for timetabling. *Discrete Mathematics*. 1997;165–166:161–170. DOI: 10.1016/S0012-365X(96)00208-7.

24. de Werra D. On a multiconstrained model for chromatic scheduling. *Discrete Applied Mathematics.* 1999;94(1–3):171–180. DOI: 10.1016/S0166-218X(99)00019-0.

25. Sotskov YN. Mixed graph colorings: a historical review. *Mathematics.* 2020;8(3):385. DOI: 10.3390/math8030385.

26. Harary F. *Graph theory*. Massachusetts: Addison-Wesley; 1969. 274 p.

27. Thulasiraman K, Swamy MNS. *Graphs: theory and algorithms*. Canada: John Wiley & Sons, Inc.; 1992. 480 p. DOI: 10.1002/9781118033104.

28. Tanaev VS, Sotskov YN, Strusevich VA. *Scheduling theory. Multi-stage systems.* Dordrecht: Kluwer Academic Publishers; 1994. 406 p. DOI: 10.1007/978-94-011-1192-8.

29. Brucker P. *Scheduling algorithms.* Berlin: Springer; 1995. 326 p. DOI: 10.1007/978-3-662-03088-2.

30. Graham RL, Lawler EL, Lenstra JK, Rinnooy Kan AHG. Optimization and approximation in deterministic sequencing and scheduling: a survey. *Annals of Discrete Mathematics.* 1979;5:287–326. DOI: 10.1016/S0167-5060(08)70356-X.

31. Sotskov YN, Tanaev VS. Scheduling theory and practice: Minsk group results. *Intelligent Systems Engineering.* 1994;3(1):1–8. DOI: 10.1049/ISE.1994.0001.

32. Sussmann B. Scheduling problems with interval disjunctions. *Mathematical Methods of Operations Research.* 1972;16:165–178.

33. Brucker P, Krämer A. Shop scheduling problems with multiprocessor tasks on dedicated processors. *Annals of Operations Research.* 1995;57(1):13–27. DOI: 10.1007/BF02099688.

34. Brucker P, Krämer A. Polynomial algorithms for resource-constrained and multiprocessor task scheduling problems. *European Journal of Operational Research.* 1996;90(2):214–226. DOI: 10.1016/0377-2217(95)00350-9.

35. Hoogeveen JA, van de Velde SL, Veltman B. Complexity of scheduling multiprocessor tasks with prespecified processor allocations. *Discrete Applied Mathematics.* 1994;55(3):259–272. DOI: 10.1016/0166-218X(94)90012-4.

36. Hoogeveen JA, Lenstra JK, Veltman B. Preemptive scheduling in a two-stage multiprocessor flow shop is NP-hard. *European Journal of Operational Research.* 1996;89(1):172–175. DOI: 10.1016/S0377-2217(96)90070-3.

Received 21.03.2021 / revised 27.06.2021 / accepted 27.06.2021.

Вычислительная математика

Computational mathematics

УДК 519.62

СТАБИЛИЗИРОВАННЫЕ ЯВНЫЕ МЕТОДЫ ТИПА АДАМСА

В. И. РЕПНИКОВ[1)]***, Б. В. ФАЛЕЙЧИК***[1)]***, А. В. МОЙСА***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Представлены явные многошаговые методы типа Адамса с расширенным интервалом устойчивости, аналогичные явным стабилизированным чебышевским методам типа Рунге – Кутты. Доказано, что для любого $k \geq 1$ существует явный k-шаговый метод типа Адамса первого порядка с интервалом устойчивости длиной $2k$. Коэффициенты и константа погрешности таких методов имеют весьма простой вид. Получена также демпфированная модификация этих методов. В общем случае для построения k-шагового метода порядка p необходимо решить задачу условной оптимизации, в которой целевая функция и p ограничений являются многочленами второй степени от k переменных. Численно построены методы до шестого порядка включительно, проведены несколько вычислительных экспериментов для подтверждения свойств аппроксимации и устойчивости.

Ключевые слова: численное решение ОДУ; жесткость; интервал устойчивости; абсолютная устойчивость; многошаговые методы; методы типа Адамса; явные методы.

Благодарность. Работа выполнена при поддержке государственной программы научных исследований Республики Беларусь «Конвергенция-2020». Авторы также выражают благодарность рецензенту статьи за подробный и компетентный отзыв.

Образец цитирования:
Репников ВИ, Фалейчик БВ, Мойса АВ. Стабилизированные явные методы типа Адамса. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:82–98 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-82-98

For citation:
Repnikov VI, Faleichik BV, Moisa AV. Stabilised explicit Adams-type methods. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:82–98.
https://doi.org/10.33581/2520-6508-2021-2-82-98

Авторы:
Василий Иванович Репников – кандидат физико-математических наук; заведующий кафедрой вычислительной математики факультета прикладной математики и информатики.
Борис Викторович Фалейчик – кандидат физико-математических наук; доцент кафедры вычислительной математики факультета прикладной математики и информатики.
Андрей Владимирович Мойса – аспирант кафедры вычислительной математики факультета прикладной математики и информатики. Научный руководитель – Б. В. Фалейчик.

Authors:
Vasily I. Repnikov, PhD (physics and mathematics); head of the department of computational mathematics, faculty of applied mathematics and computer science.
repnikov@bsu.by
Boris V. Faleichik, PhD (physics and mathematics); associate professor at the department of computational mathematics, faculty of applied mathematics and computer science.
faleichik@bsu.by
Andrew V. Moisa, postgraduate student at the department of computational mathematics, faculty of applied mathematics and computer science.
moysa@bsu.by

STABILISED EXPLICIT ADAMS-TYPE METHODS

V. I. REPNIKOV [a], ***B. V. FALEICHIK*** [a], ***A. V. MOISA*** [a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: B. V. Faleichik (faleichik@bsu.by)

In this work we present explicit Adams-type multi-step methods with extended stability intervals, which are analogous to the stabilised Chebyshev Runge – Kutta methods. It is proved that for any $k \geq 1$ there exists an explicit k-step Adams-type method of order one with stability interval of length $2k$. The first order methods have remarkably simple expressions for their coefficients and error constant. A damped modification of these methods is derived. In the general case, to construct a k-step method of order p it is necessary to solve a constrained optimisation problem in which the objective function and p constraints are second degree polynomials in k variables. We calculate higher-order methods up to order six numerically and perform some numerical experiments to confirm the accuracy and stability of the methods.

Keywords: numerical ODE solution; stiffness; stability interval; absolute stability; multi-step methods; Adams-type methods; explicit methods.

Acknowledgements. The work is supported by Belarusian government program of scientific research «Convergence-2020». The authors also would like to thank the anonymous reviewer for valuable comments and suggestions.

Introduction

A k-step explicit Adams-type method for the numerical integration of the ordinary differential equation system

$$y' = f(t, y),\ y(t_0) = y_0,\ y : \mathbb{R} \to \mathbb{R}^n,\ f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n,$$

on a uniform grid has the form

$$y_{m+k} = y_{m+k-1} + \tau\left(\beta_0 f_m + \ldots + \beta_{k-1} f_{m+k-1}\right), \tag{1}$$

where m is the step number, $y_l \approx y(t_0 + l\tau)$, $f_l = f(t_0 + l\tau, y_l)$, $l \geq 0$, τ is a discretisation step, and $\{\beta_j\}$, $j = 0, 1, \ldots, k-1$, are the coefficients for the method.

A conventional means of analysing the linear stability of a multi-step method is through the construction of a stability region $S \subset \mathbb{C}$ such that for all $\lambda\tau \in S$ the numerical solution of the model linear problem

$$y' = \lambda y,\ \lambda \in \mathbb{C},$$

remains bounded for all m. The equivalent requirement is that all roots ζ_j of the characteristic equation

$$\rho(\zeta) - \lambda\tau\sigma(\zeta) = 0 \tag{2}$$

lie within the unit disc on the complex plane, and all the roots of modulus one are simple [1]. Here ρ and σ are the standard generating polynomials which in our case have the form

$$\rho(\zeta) = \zeta^k - \zeta^{k-1},\ \sigma(\zeta) = \sum_{j=0}^{k-1} \beta_j \zeta^j.$$

The stability interval of a method is the largest interval of the real axis of the form $[-\ell, 0]$ contained in S. Here $\ell \geq 0$ is the value which will be referred to as the length of the stability interval. As is known, stability intervals of the classical explicit Adams methods are small and get smaller with the growth of k, so these methods are not suitable for stiff problems. The purpose of the present research is to construct explicit multi-step methods of Adams type (1) of order $p < k$ with increased lengths of stability intervals. Putting it in other words, we develop a multi-step analog of the well-known Chebyshev Runge – Kutta methods [2–5].

The main obstacle for the way of construction of multi-step methods for stiff problems is the dependence of the error constant on the size of stability region, which was investigated by Jeltsch and Nevanlinna [6; 7]. On the one hand, due to [5, theorem 4.2], for any $k > 1$, $\alpha < \frac{\pi}{2}$ and $R > 0$ there exists an explicit linear multi-step method of order $k - 1$ such that the method is stable in the set

$$\{\mu \in \mathbb{C} : |\mu| \leq R, |\arg(-\mu)| \leq \alpha\}.$$

Unfortunately, methods which stability regions contain large disks of the form $\{\mu \in \mathbb{C} : |\mu + R| \leq R\}$ are useless in practice due to huge error constants (norms of Peano kernels) [6, theorem 4.1; 1, chapter V.2, theorem 2.6].

On the other hand, in the case of long stability intervals the lower bounds for the error constant are less restrictive. Namely [6, theorem 4.4] gives a lower bound for Peano kernel norms for explicit multi-step methods in the form of $C_k\ell$, where $C_k > 0$. Another interesting fact from [6, theorem 4.7] is that for explicit k-step methods of order $k-1$ with $\ell > 2$ the error constant has lower bound equal to $\delta_{k-1}\dfrac{\ell-2}{2^{k-1}}$, where (δ_k) is a decreasing sequence with $\delta_1 = 0.5$. Thus, we hope that explicit multi-step methods with extended stability interval can have reasonable error constants (see section «Higher order methods»).

The material is organised as follows. In section «Optimisation strategy» we describe the general framework of the method's construction, sections «First order methods» and «First order methods with damping» are devoted to the first order methods and their damped modifications. Higher order methods are discussed in section «Higher order methods», and section «Numerical experiments» contains the results of numerical experiments. In the last section we discuss the obtained results and make final conclusions.

Optimisation strategy

The conventional way of constructing a stability region is to find a root locus curve $\mathcal{C}$ defined as

$$\mathcal{C} = \left\{\mu_\beta\left(e^{i\varphi}\right) : \varphi \in [0,\, 2\pi)\right\}, \tag{3}$$

where the function $\mu_\beta : \mathbb{C} \to \mathbb{C}$ maps a root of the characteristic equation (2) to the corresponding value of $\lambda\tau$:

$$\mu_\beta(\zeta) = \frac{\rho(\zeta)}{\sigma(\zeta)}. \tag{4}$$

The subscript β indicates the dependence on the coefficients of method (1) which we are to be determined. From the definition of stability region it follows that $\partial S \subseteq \mathcal{C}$ and

$$\ell \le \mu_\beta(-1),$$

thus the optimisation problem can be stated as

$$\beta^* = \operatorname*{argmin}_{\beta \in \mathcal{F} \cap \mathcal{P}} \mu_\beta(-1), \tag{5}$$

where $\mathcal{P}$ is a set of coefficients which satisfy the posed order conditions, and $\mathcal{F}$ is a feasible set of coefficients with a desired shape of the root locus curve. This set is defined as follows.

Primarily we would like to have $\ell = -\mu_\beta(-1)$ for all $\beta \in \mathcal{F}$. To assure this we require the locus curve (3) not to cross the real axis before the parameter φ reaches π. This condition triggers the following definition for the feasible set:

$$\mathcal{F} = \left\{\beta \in \mathbb{R}^k : \operatorname{Im} \mu_\beta\left(e^{i\varphi}\right) \ge 0 \ \forall \varphi \in (0,\, \pi)\right\}. \tag{6}$$

The main question now is how to find a parametrisation of $\mathcal{F}$ which allows the reduction of (5), (6) to some manageable form. We start by noting that

$$\operatorname{Im} \mu_\beta\left(e^{i\varphi}\right) \ge 0 \Leftrightarrow \nu_\beta(\varphi) \ge 0,$$

where

$$\nu_\beta(\varphi) = \operatorname{Im} \rho\left(e^{i\varphi}\right)\overline{\sigma\left(e^{i\varphi}\right)} = \sum_{j=1}^{k} \left(\beta_{k-j} - \beta_{k-j-1}\right)\sin j\varphi. \tag{7}$$

From here we set $\beta_j = 0$ for all $j < 0$ and $j > k-1$. By utilising the Chebyshev polynomials of the second kind U_j,

$$U_{j-1}(\cos\varphi)\sin\varphi = \sin j\varphi,$$

and using the power reduction formulae for the powers of $\cos\varphi$, (7) can be represented as

$$\nu_\beta(\varphi) = \sin\varphi \sum_{j=0}^{k-1} a_j \cos j\varphi \tag{8}$$

with some $a_j \in \mathbb{R}$. Since we need $\nu_\beta(\varphi)$ to be non-negative on $(0,\, \pi)$, the following result from [7, lemma 6.1.3] is useful.

Lemma. *For any non-negative trigonometric polynomial a of the form*

$$A(\varphi) = \sum_{j=0}^{k} a_j \cos j\varphi,\ a_j \in \mathbb{R},$$

there exists a trigonometric polynomial

$$B(\varphi)=\sum_{j=0}^{k} b_j e^{ij\varphi},\ b_j\in\mathbb{R},$$

such that $A(\varphi)=\left|B(\varphi)\right|^2$.

From this lemma it follows that all feasible trigonometric polynomials ν_β have the form

$$\nu_\beta(\varphi)=\sin\varphi\left|\sum_{j=0}^{k-1} b_j e^{ij\varphi}\right|^2=\sin\varphi\sum_{j,\,l=0}^{k-1} b_j b_l\cos(j-l)\varphi \tag{9}$$

with $b_j\in\mathbb{R}$. To complete the transformation of the optimisation problem we must express the original coefficients $\{\beta_j\}$ in terms of $\{b_j\}$. This can be done by converting (8) to the same basis of sin $j\varphi$ as (7):

$$\sin\varphi\sum_{j=0}^{k-1} a_j\cos j\varphi=\frac{1}{2}\sum_{j=0}^{k-1} a_j\left(\sin(j+1)\varphi-\sin(j-1)\varphi\right).$$

By equating this expression with (7) it is straightforward to get

$$\beta_j=\frac{1}{2}\left(\tilde a_{j-1}+\tilde a_j\right),\ j=0,\,1,\,\ldots,\,k-2, \tag{10}$$

$$\beta_{k-1}=\tilde a_{k-1}+\frac{\tilde a_{k-2}}{2}. \tag{10a}$$

Here for clarity $\tilde a_j=a_{k-1-j}$, $\tilde a_{-1}=0$. Conversely, from (8), (9) we have

$$\tilde a_j=2\sum_{l=0}^{j} b_l b_{k-1+l-j},\ j=0,\,1,\,\ldots,\,k-2, \tag{10b}$$

$$\tilde a_{k-1}=\sum_{j=0}^{k-1} b_j^2. \tag{10c}$$

Transformations (10)–(10c) define the required mapping

$$T: b\to\beta,$$

where $b=\left(b_0,\,\ldots,\,b_{k-1}\right)$, $\beta=\left(\beta_0,\,\ldots,\,\beta_{k-1}\right)$.

Now let us derive the form of objective function (5),

$$\mu_\beta(-1)=\frac{2\cdot(-1)^k}{\sum_{j=0}^{k-1}(-1)^j\beta_j}, \tag{11}$$

in terms of b. By direct application of (10)–(10c) we have

$$\sum_{j=0}^{k-1}(-1)^j\beta_j=(-1)^k\sum_{j=0}^{k-1} b_j^2,$$

so the initial optimisation problem (5), (6) finally takes the surprisingly simple form

$$b^*=\operatorname*{argmin}_{\beta\in\mathcal{P}_p'}\sum_{j=0}^{k-1} b_j^2, \tag{12}$$

where $\mathcal{P}_p'$ is the order p restriction set:

$$\mathcal{P}_p'=\left\{b\in\mathbb{R}^k: G_q\left(T(b)\right)=0,\ q=1,\,2,\,\ldots,\,p\right\}, \tag{12a}$$

$$G_1(\beta)=\sum_{j=0}^{k-1}\beta_j-1, \tag{12b}$$

$$G_q(\beta)=\sum_{j=0}^{k-1}(1-k+j)^{q-1}\beta_j-\frac{1}{q},\ q>1. \tag{12c}$$

First order methods

Theorem 1. *For any number of steps $k > 1$ there exists a first order explicit Adams-type method with stability interval of length $2k$. The method has the form*

$$y_{k+m} = y_{k+m-1} + \frac{\tau}{k^2}\left(f_m + 3f_{m+1} + \ldots + (2k-1)f_{m+k-1}\right), \tag{13}$$

i. e., $\beta_j = \dfrac{2j+1}{k^2}$.

P r o o f. Directly applying (10)–(10c) to the first order condition (12b) we have

$$\beta_0 + \beta_1 + \ldots + \beta_{k-1} = a_0 + a_1 + \ldots + a_{k-1} = 1.$$

Conversely, by the construction form (8), (9) we have

$$\sum_{j=0}^{k-1} a_j = \sum_{j,\,l}^{k-1} b_j b_l = \left(\sum_{j=1}^{k-1} b_j\right)^2.$$

Thus, the optimisation problem (12)–(12c) in the case of $p = 1$ takes the form

$$b^* = \operatorname*{argmin}_{b \in \mathbb{R}^k}\left(b_0^2 + b_1^2 + \ldots + b_{k-1}^2\right),$$

subject to

$$\left(b_0 + b_1 + \ldots + b_{k-1}\right)^2 = 1.$$

Solving this problem by the method of Lagrange multipliers, we get $b_j^* = k^{-1}$ for all j and then by (10)–(10c) obtain

$$\beta^* = T\left(b^*\right) = \left(\frac{1}{k^2}, \frac{3}{k^2}, \ldots, \frac{2k-1}{k^2}\right).$$

By construction of the method from (11) and since

$$\sum_{j=0}^{k-1} (-1)^j \frac{2j+1}{k^2} = (-1)^k \frac{1}{k}, \tag{14}$$

we have $\ell = -\mu_{\beta^*}(-1) = 2k$.

There is an interesting parity of the above result with the case of s-stage Chebyshev Runge – Kutta methods of order 1, which require s evaluations of f per step and have stability interval equal to $2s^2$ [1; 2]. This allows us to suppose that the achieved length of $2k$ is the largest possible for explicit first order multi-step methods.

Error constant. According to [9] we define the error constant of the multi-step method as

$$C = \frac{C_{p+1}}{\sigma(1)},$$

where

$$C_{p+1} = \frac{1}{(p+1)!}\sum_{j=0}^{k}\left(a_j j^{p+1} - (p+1)\beta_j j^p\right).$$

It is easy to calculate this constant in our case.

Proposition 1. *The error constant of the optimised first order methods is equal to*

$$C = \frac{k}{3} + \frac{1}{6k}.$$

P r o o f. Since $\beta_j = \dfrac{2j+1}{k^2}$, $\alpha_k = 1$, $\alpha_{k-1} = -1$, $\sigma(1) = 1$, we have

$$C = \frac{1}{2}\left(k^2 - (k-1)^2 - \frac{2}{k^2}\sum_{j=0}^{k} j(2j+1)\right) = \frac{k}{3} + \frac{1}{6k}.$$

First order methods with damping

Analogously to the Chebyshev Runge – Kutta methods, in order to pull the root locus curve away from the real axis for $\varphi \in (0, \pi)$ it is necessary to perform a damping transformation with the constructed methods.

Using (7), (8) consider (4) and represent

$$\operatorname{Im}\mu_\beta\left(e^{i\varphi}\right) = Q(\varphi)\sin\varphi,$$

where

$$Q(\varphi) = \frac{\sum\limits_{j=0}^{k-1} a_j \cos j\varphi}{\left|\sigma\left(e^{i\varphi}\right)\right|^2} = \frac{\sum\limits_{j=0}^{k-1} a_j \cos j\varphi}{\sum\limits_{j=0}^{k-1} \delta_j \cos j\varphi}, \tag{15}$$

and

$$\delta_0 = \sum_{j=0}^{k-1} \beta_j^2,\ \delta_j = 2\sum_{l=0}^{j} \beta_l \beta_{j+l}. \tag{16}$$

Recall that the connection between coefficients a_j and β_j is described by the first two equalities of (10)–(10c). Let $\hat{Q}(\varphi)$ be the damped method's counterpart of (15). We define it as

$$\hat{Q}(\varphi) = \frac{\sum\limits_{j=0}^{k-1} \hat{a}_j \cos j\varphi}{\sum\limits_{j=0}^{k-1} \hat{\delta}_j \cos j\varphi} = C\left(Q(\varphi) + \varepsilon\right),$$

where ε controls the «shift» from the real axis and C is a scaling constant to be determined. Then we have $\hat{a}_j = Ca'_j$,

$$a'_j = a_j + \varepsilon\delta_j.$$

Now we use this equality together with (10)–(10c) and get

$$\beta'_j = \beta_j + \frac{\varepsilon}{2}\left(\delta_{k-j} + \delta_{k-j-1}\right),\ j = 0, 1, \ldots, k-2,$$

$$\beta'_{k-1} = \beta_{k-1} + \frac{\varepsilon}{2}\delta_1 + \varepsilon\delta_0,\ \delta_k = 0.$$

The coefficients $\hat{\beta}_j$ of the sought damped method are expressed as $\hat{\beta}_j = C\beta'_j$. To keep the order of the method equal to one, the constant C should be equal to $\left(\sum\limits_{j=0}^{k-1} \beta'_j\right)^{-1}$. By (16) we have

$$\sum_{j=0}^{k-1} \beta'_j = \sum_{j=0}^{k-1} \beta_j + \varepsilon\sum_{j=0}^{k-1} \delta_j = \sum_{j=0}^{k-1} \beta_j + \varepsilon \sum_{j,\,l=0}^{k-1} \beta_j\beta_l.$$

Since $\sum\limits_{j=0}^{k-1} \beta_j = 1$ we finally obtain the following formulae for the coefficients of the damped method:

$$\hat{\beta}_j = \frac{\beta_j + \varepsilon\Delta_j}{1+\varepsilon},\ j = 0, 1, \ldots, k-1, \tag{17}$$

where

$$\Delta_j = \frac{1}{2}\left(\delta_{k-j} + \delta_{k-j-1}\right),\ j = 0, 1, \ldots, k-2, \tag{17a}$$

$$\Delta_{k-1} = \frac{1}{2}\delta_1 + \delta_0. \tag{17b}$$

The values of Δ_j for k from 2 to 10 for the optimised first order method (13) are presented in table 1. The stability region boundaries of the one-step methods together with their damped counterparts are displayed in fig. 1.

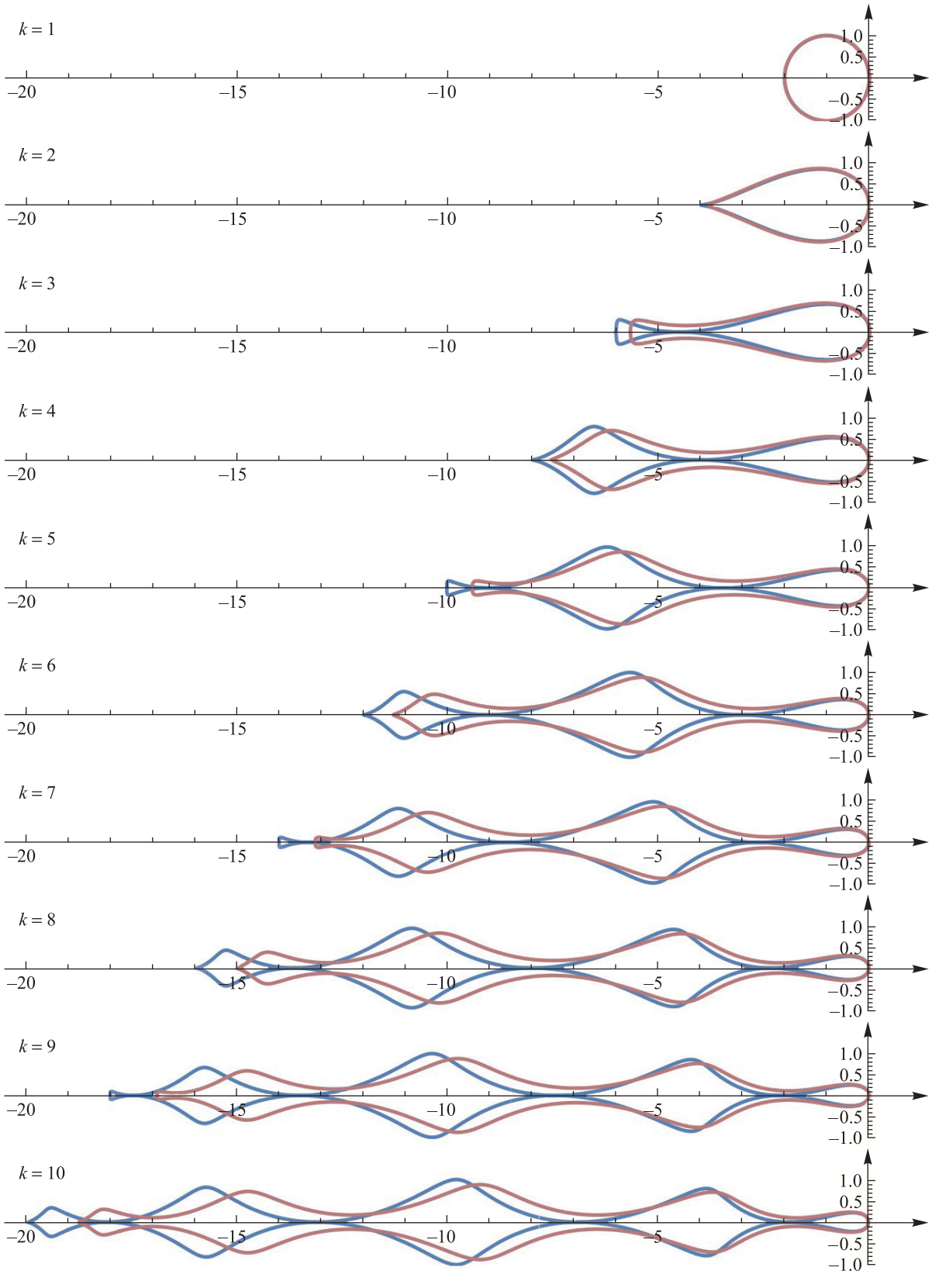

Fig. 1. Stability regions of the optimised first order methods and their damped versions (ε = 0.25) for $k = 1, \ldots, 10$

Table 1

The values of Δ_j for determining the coefficients of the optimised first order methods with damping (17), $\beta_j = \frac{2j+1}{k^2}$

k	Δ_0	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9
2	$\frac{3}{16}$	$\frac{13}{16}$								
3	$\frac{5}{81}$	$\frac{23}{81}$	$\frac{53}{81}$							
4	$\frac{7}{256}$	$\frac{33}{256}$	$\frac{79}{256}$	$\frac{137}{256}$						
5	$\frac{9}{625}$	$\frac{43}{625}$	$\frac{21}{625}$	$\frac{187}{625}$	$\frac{281}{625}$					
6	$\frac{11}{1296}$	$\frac{53}{1296}$	$\frac{131}{1296}$	$\frac{79}{1296}$	$\frac{121}{1296}$	$\frac{167}{1296}$				
7	$\frac{13}{2401}$	$\frac{9}{343}$	$\frac{157}{2401}$	$\frac{41}{343}$	$\frac{445}{2401}$	$\frac{89}{343}$	$\frac{813}{2401}$			
8	$\frac{15}{4096}$	$\frac{73}{4096}$	$\frac{183}{4096}$	$\frac{337}{4096}$	$\frac{527}{4096}$	$\frac{745}{4096}$	$\frac{983}{4096}$	$\frac{1233}{4096}$		
9	$\frac{17}{6561}$	$\frac{83}{6561}$	$\frac{209}{6561}$	$\frac{43}{729}$	$\frac{203}{2187}$	$\frac{289}{2187}$	$\frac{1153}{6561}$	$\frac{1459}{6561}$	$\frac{1777}{6561}$	
10	$\frac{19}{10\,000}$	$\frac{93}{10\,000}$	$\frac{47}{2000}$	$\frac{437}{10\,000}$	$\frac{691}{10\,000}$	$\frac{989}{10\,000}$	$\frac{1323}{10\,000}$	$\frac{337}{2000}$	$\frac{2067}{10\,000}$	$\frac{2461}{10\,000}$

Proposition 2. *The stability interval of the damped method (17)–(17b) with* $\beta_j = \beta_j^* = \frac{2j+1}{k^2}$ *is equal to* $\left[-\ell_\varepsilon, 0\right]$, *where*

$$\ell_\varepsilon = \frac{6(1+\varepsilon)k^3}{\varepsilon\left(4k^2-1\right)+3k^2}.$$

P r o o f. Let us compute

$$\sum_{j=0}^{k-1}(-1)^j\hat{\beta}_j = (1+\varepsilon)^{-1}\left(\sum_{j=0}^{k-1}(-1)^j\beta_j^* + \varepsilon(-1)^{k-1}\delta_0^*\right).$$

The first term has already been calculated in (14) and the second is determined as

$$\delta_0^* = \sum_{j=0}^{k-1}\left(\beta_j^*\right)^2 = \sum_{j=0}^{k-1}\frac{(2j+1)^2}{k^4} = \frac{4k^2-1}{3k^3}.$$

From here we finally get

$$\mu_{\hat{\beta}}(-1) = \frac{2(-1)^k}{\sum\limits_{j=0}^{k-1}(-1)^j\hat{\beta}_j} = -\frac{6(1+\varepsilon)k^3}{\varepsilon\left(4k^2-1\right)+3k^2}.$$

Corollary 1. *The asymptotic length of the damped one-step method is*

$$\lim_{\varepsilon\to\infty}\ell_\varepsilon = \frac{3}{2}k.$$

Higher order methods

To construct a stabilised k-step Adams-type method of order p one should use the general form of the optimisation problem (12)–(12c) with mapping T specified by (10)–(10c). For example, for $k = 5$, $p = 4$, the problem in terms of b_j takes the form

minimise

$$b_0^2 + b_1^2 + b_2^2 + b_3^2 + b_4^2$$

subject to

$$\left(b_0 + b_1 + b_2 + b_3 + b_4\right)^2 = 1,$$

$$b_2b_3 + \left(3b_2 + b_3\right)b_4 + b_1\left(b_2 + 3b_3 + 5b_4\right) + b_0\left(b_1 + 3b_2 + 5b_3 + 7b_4\right) = -\frac{1}{2},$$

$$b_2b_3 + \left(5b_2 + b_3\right)b_4 + b_1\left(b_2 + 5b_3 + 13b_4\right) + b_0\left(b_1 + 5b_2 + 13b_3 + 25b_4\right) = \frac{1}{3},$$

$$b_2b_3 + \left(9b_2 + b_3\right)b_4 + b_1\left(b_2 + 9b_3 + 35b_4\right) + b_0\left(b_1 + 9b_2 + 35b_3 + 91b_4\right) = -\frac{1}{4}.$$

The symbolic solution of this problem yielded by *Wolfram Mathematica* after transforming back to the initial variables β_j is

$$\beta^* = \left(-\frac{1}{4}, \frac{5}{8}, \frac{1}{24}, -\frac{35}{24}, \frac{49}{24}\right)$$

with $\ell = 0.75$, to compare with $\ell = 0.3$ for the classical explicit Adams-type method of order 4. Another neat example is the 5-step method of order 2:

$$\beta^* = \left(-\frac{1}{8}\left(3 - \sqrt{5}\right), -\frac{3}{4}\left(\sqrt{5} - 2\right), 0, \frac{7}{4}\left(\sqrt{5} - 2\right), \frac{9}{8}\left(3 - \sqrt{5}\right)\right)$$

with $\ell = 2 + \frac{4}{\sqrt{5}} \approx 3.789$. The stability regions of these and the rest of the 5-step methods are shown in the uppermost part of fig. 2.

Unfortunately, it is not always possible to obtain the solution symbolically, thus we compute the coefficients of our methods numerically using *Mathematica's* function `NMinimise`, see the corresponding code in Appendix A. We used 50-digit working precision and computed the (k, p) methods for k from 3 to 10 and p from 2 to k (with the latter value corresponding to the classical Adams methods). The results with 20-digit accuracy are displayed in tables 3 and 4. It is interesting that the (4, 3) method coincides with [9, formula (5.4)]. The stability regions of 5-, 6- and 7-step methods are shown in fig. 2.

To assess the accuracy of the obtained coefficients we checked the magnitude of the order residuals $G_q\left(\beta^*\right)$ see (12b), (12c). In all convergent cases these residuals do not exceed 10^{-19}. Note that *Mathematica's* function `NMinimise` failed to converge in the following cases: $(k, p) = (7, 6)$ and for all $p > 7$. Our hypothesis is that for these (k, p) combinations the Adams-type methods satisfying our restrictions do not exist. Note that (11, 7) method seems to exist and has microscopic $\ell \approx 0.051$ which is just slightly more than $\ell = 0.0465$ for the (7, 7) case. The error constants of all the calculated methods are presented in table 2.

Table 2

Error constants of the stabilised Adams-type methods

k	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$
2	0.75					
3	1.0556	0.66667				
4	1.375	1.0380	0.62500			
5	1.7	1.5208	1.0227	0.59861		
6	2.0278	2.1128	1.5972	1.0120	0.57928	
7	2.3571	2.8134	2.3814	1.6471	1.0032	
8	2.6875	3.6223	3.4092	2.5751	1.6825	0.99505
9	3.0185	4.5392	4.7148	3.8788	2.7235	1.7079
10	3.35	5.5643	6.3328	5.6524	4.2616	2.8403

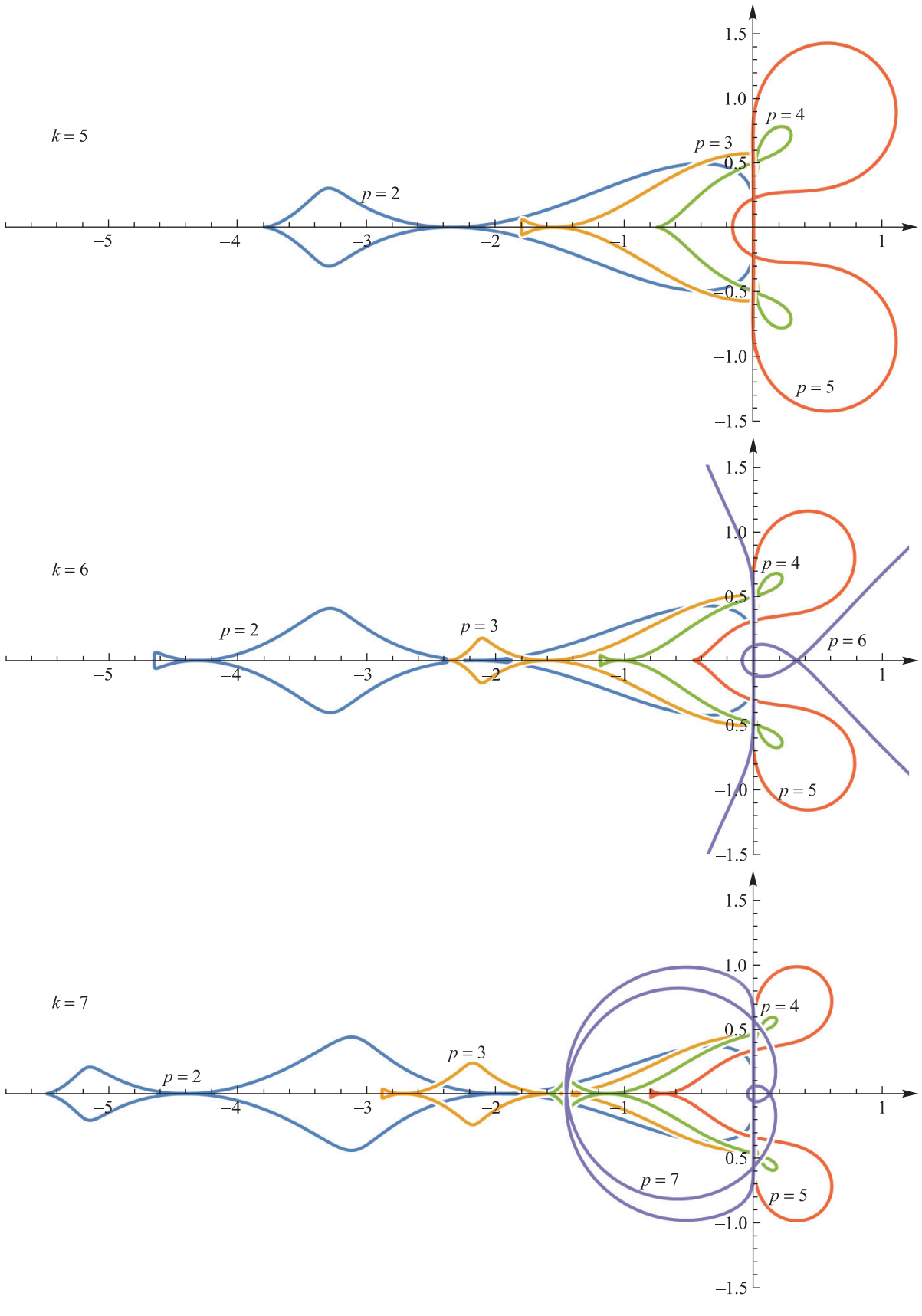

Fig. 2. Stability regions of the multi-step optimised methods for $k = 5, 6, 7$

Coefficients and stability interval lengths

k	Order 2	Order 3
3	$\ell = 2$ −0.25 0 1.25	$\ell = 0.545\,454\,545\,454\,545\,455$ 0.416 666 666 666 666 666 67 −1.333 333 333 333 333 333 3 1.916 666 666 666 666 666 7
4	$\ell = 2.914\,213\,562\,373\,095$ −0.146 446 609 406 730 460 69 −0.181 980 515 339 459 636 91 0.303 300 858 899 110 655 90 1.025 126 265 847 079 441 7	$\ell = 1.2$ 0.25 −0.333 333 333 333 333 333 33 −0.583 333 333 333 333 333 33 1.666 666 666 666 666 666 7
5	$\ell = 3.788\,854\,381\,999\,832$ −0.095 491 502 812 526 287 949 −0.177 050 983 124 842 272 31 0 0.413 118 960 624 631 968 72 0.859 423 525 312 736 591 54	$\ell = 1.793\,779\,334\,348\,686$ 0.164 376 941 012 461 256 19 −0.009 791 091 775 013 643 927 1 −0.540 221 704 083 059 938 67 −0.047 691 080 558 684 215 630 1.433 326 935 404 296 542 0
6	$\ell = 4.642\,734\,410\,091\,836$ −0.066 987 298 107 786 995 665 −0.147 114 317 029 978 077 15 −0.089 745 962 155 603 046 598 0.125 644 347 017 869 431 07 0.441 342 951 089 917 564 59 0.736 860 279 185 581 123 75	$\ell = 2.347\,826\,086\,956\,522$ 0.115 740 740 740 747 316 06 0.087 962 962 962 956 387 640 −0.287 037 037 037 050 187 68 −0.407 407 407 407 394 256 76 0.245 370 370 370 376 945 69 1.245 370 370 370 363 795 0
7	$\ell = 5.484\,476\,959\,454\,063$ −0.049 515 566 048 790 436 882 −0.119 125 202 772 785 772 27 −0.110 182 500 025 524 205 85 0 0.198 328 500 045 943 570 54 0.436 792 410 166 881 165 01 0.643 702 358 634 275 679 47	$\ell = 2.877\,558\,710\,633\,067$ 0.085 721 156 820 309 456 282 0.111 546 128 114 633 279 41 −0.117 210 331 348 086 366 45 −0.354 636 657 791 245 849 07 −0.217 440 005 322 052 225 76 0.395 573 727 915 164 329 79 1.096 445 981 611 277 375 8
8	$\ell = 6.318\,535\,592\,272\,045$ −0.038 060 233 744 366 798 686 −0.096 797 724 520 983 369 102 −0.107 796 952 873 510 886 96 −0.052 994 558 379 770 972 895 0.068 135 860 774 038 963 863 0.237 153 296 321 735 328 73 0.419 456 806 257 518 582 77 0.570 903 506 165 339 152 29	$\ell = 3.391\,689\,975\,797\,208$ 0.065 966 021 983 597 280 828 0.110 324 410 873 232 080 03 −0.022 713 554 363 876 414 313 −0.236 910 211 826 378 789 68 −0.306 342 255 793 388 331 74 −0.055 862 376 110 155 554 778 0.468 251 519 600 799 889 06 0.977 286 445 636 169 840 59
9	$\ell = 7.147\,430\,550\,561\,413$ −0.030 153 689 607 037 932 268 −0.079 550 128 858 107 345 641 −0.098 407 115 533 249 091 604 −0.073 305 865 502 781 992 742 0 0.115 194 931 504 337 426 25 0.255 858 500 386 456 032 91 0.397 750 644 290 616 707 00 0.512 612 723 319 786 611 24	$\ell = 3.895\,290\,219\,607\,647$ 0.052 301 051 895 272 605 013 0.101 266 962 101 188 747 90 0.026 642 016 446 313 140 531 −0.137 931 929 156 682 986 04 −0.266 954 905 132 604 002 00 −0.219 076 590 372 349 287 46 0.064 279 256 258 874 647 289 0.499 021 276 364 748 587 44 0.880 452 861 595 238 547 33
10	$\ell = 7.972\,691\,637\,812\,280$ −0.024 471 741 852 422 821 505 −0.066 228 831 765 768 206 903 −0.087 599 164 129 385 382 526 −0.078 738 975 641 538 713 579 −0.034 883 488 233 566 344 682 0.042 635 374 507 685 291 073 0.146 229 526 191 426 841 03 0.262 797 492 388 163 164 20 0.375 296 713 339 364 715 57 0.464 963 095 196 041 457 33	$\ell = 4.391\,469\,108\,714\,782$ 0.042 467 110 956 300 544 552 0.090 440 497 652 067 647 206 0.051 030 056 647 860 180 918 −0.068 250 050 077 061 163 328 −0.199 020 948 512 629 179 34 −0.243 955 170 425 047 826 18 −0.129 131 045 380 918 158 96 0.148 969 943 352 139 819 08 0.506 940 597 805 911 675 08 0.800 509 007 981 376 460 97

Table 3

of the optimised methods of order 2–5

Order 4	Order 5
$\ell = 0.3$ −0.375 000 000 000 000 000 00 1.541 666 666 666 666 666 7 −2.458 333 333 333 333 333 3 2.291 666 666 666 666 666 7	
$\ell = 0.75$ −0.25 0.625 0.041 666 666 666 666 666 667 −1.458 333 333 333 333 333 3 2.041 666 666 666 666 666 7	$\ell = 0.163\,339\,382\,940\,108\,8$ 0.348 611 111 111 111 111 11 −1.769 444 444 444 444 444 4 3.633 333 333 333 333 333 3 −3.852 777 777 777 777 777 8 2.640 277 777 777 777 777 8
$\ell = 1.181\,897\,711\,989\,360$ −0.176 228 059 145 769 668 84 0.217 779 624 303 801 742 76 0.516 162 094 242 489 717 34 −0.676 216 770 425 916 253 54 −0.686 030 943 361 995 271 80 1.804 534 054 387 389 734 1	$\ell = 0.469\,157\,254\,561\,251$ 0.249 421 296 296 296 296 29 −0.898 495 370 370 370 370 36 0.724 768 518 518 518 518 51 1.139 120 370 370 370 370 4 −2.605 671 296 296 296 296 3 2.390 856 481 481 481 481 5
$\ell = 1.586\,803\,103\,995\,642$ −0.130 276 570 699 249 748 82 0.040 823 321 662 060 514 133 0.451 574 102 013 991 105 94 0.016 001 789 308 425 411 316 −0.794 867 969 474 415 111 95 −0.187 023 021 147 214 511 56 1.603 768 348 336 402 340 9	$\ell = 0.792\,362\,028\,995\,767$ 0.184 805 705 228 950 410 08 −0.435 462 017 690 876 205 65 −0.246 164 378 868 764 011 81 1.268 163 587 804 809 902 2 −0.328 303 225 060 673 063 70 −1.594 750 940 737 348 964 2 2.151 711 269 323 901 933 0
$\ell = 1.970\,916\,561\,391\,601$ −0.100 018 782 547 822 773 31 −0.035 748 890 463 804 216 949 0.306 583 717 661 130 874 97 0.277 490 855 549 241 809 02 −0.335 165 400 358 394 580 87 −0.671 991 651 703 567 700 75 0.121 222 314 597 282 248 06 1.437 627 837 265 934 339 8	$\ell = 1.105\,498\,503\,602\,666$ 0.141 608 310 782 164 335 00 −0.194 778 897 037 351 300 08 −0.452 472 522 388 392 286 71 0.576 366 301 237 590 321 03 0.783 547 082 304 835 859 81 −0.919 538 234 654 641 505 58 −0.877 252 163 864 666 963 27 1.942 520 123 620 461 539 7
$\ell = 2.339\,983\,407\,348\,191$ −0.079 129 092 227 346 338 565 −0.067 460 438 055 823 679 907 0.185 229 899 631 699 256 08 0.316 756 417 686 937 500 27 0.007 699 688 785 598 755 599 3 −0.485 616 427 960 531 390 31 −0.486 411 071 972 200 782 01 0.308 966 990 668 254 142 62 1.299 964 033 443 412 536 2	$\ell = 1.405\,151\,117\,615\,213$ 0.111 679 587 452 253 674 79 −0.068 703 909 200 215 014 827 −0.415 591 348 832 787 799 76 0.075 957 984 853 647 800 951 0.789 757 119 684 457 386 45 0.168 798 574 062 768 170 77 −1.038 227 745 131 630 760 2 −0.387 719 877 150 909 038 34 1.764 049 614 262 415 580 2
$\ell = 2.698\,087\,099\,023\,256$ −0.064 133 502 960 306 610 717 −0.078 573 353 260 495 406 661 0.099 782 736 471 490 155 539 0.274 091 499 569 754 023 55 0.175 219 063 810 426 583 79 −0.202 657 937 197 901 007 91 −0.503 462 625 956 399 647 88 −0.307 138 431 963 688 427 39 0.421 961 371 543 814 430 77 1.184 911 179 943 305 906 9	$\ell = 1.692\,885\,048\,664\,239$ 0.090 219 510 737 302 839 601 −0.002 158 456 205 061 795 703 7 −0.321 954 875 526 057 453 95 −0.171 484 785 692 822 685 95 0.474 867 894 821 556 848 85 0.598 397 647 261 845 953 95 −0.276 718 534 444 465 663 97 −0.946 384 003 148 205 677 30 −0.057 121 557 681 252 610 888 1.612 337 159 877 160 245 3

Coefficients and stability interval lengths

k	Order 6	Order 7
6	ℓ = 0.087 719 298 245 614 04 –0.329 861 111 111 111 111 11 1.997 916 666 666 666 666 7 –5.068 055 555 555 555 555 6 6.931 944 444 444 444 444 4 –5.502 083 333 333 333 333 3 2.970 138 888 888 888 888 9	
7	NOT CONVERGED	ℓ = 0.046 513 917 259 370 46 0.315 591 931 216 931 216 93 –2.223 412 698 412 698 412 7 6.731 795 634 920 634 920 6 –11.379 894 179 894 179 894 11.665 823 412 698 412 698 –7.395 634 920 634 920 634 9 3.285 730 820 105 820 105 8
8	ℓ = 0.529 072 293 477 333 5 –0.191 136 896 168 322 945 85 0.658 500 132 899 500 866 28 –0.266 987 088 973 334 445 93 –1.504 164 071 623 426 548 7 1.831 315 884 128 336 433 4 0.753 947 159 797 829 986 77 –2.763 292 764 839 035 425 9 2.481 817 644 778 452 079 9	NOT CONVERGED
9	ℓ = 0.774 504 411 366 456 2 –0.150 724 057 709 530 551 68 0.366 164 179 624 831 523 68 0.264 862 407 421 359 273 31 –1.167 936 096 027 017 846 0 –0.049 706 767 276 153 478 305 1.830 740 825 812 214 481 7 –0.540 064 514 417 873 107 85 –1.820 117 139 586 925 971 6 2.266 781 162 159 095 680 2	NOT CONVERGED
10	ℓ = 1.015 322 150 308 401 –0.121 499 259 815 889 551 61 0.195 020 012 105 151 545 22 0.403 236 549 673 635 503 99 –0.602 004 140 817 800 156 59 –0.798 017 750 437 058 784 58 0.912 988 626 427 640 081 11 1.164 843 723 085 023 816 7 –1.100 111 173 235 220 067 2 –1.133 472 337 616 751 702 8 2.079 015 750 631 269 315 8	NOT CONVERGED

Table 4

of the optimised methods of order 6–9

Order 8	Order 9
$\ell = 0.024\,408\,513\,276\,164\,89$ $-0.304\,224\,537\,037\,037\,037\,04$ $2.445\,163\,690\,476\,190\,476\,2$ $-8.612\,127\,976\,190\,476\,190\,5$ $17.379\,654\,431\,216\,931\,217$ $-22.027\,752\,976\,190\,476\,190$ $18.054\,538\,690\,476\,190\,476$ $-9.525\,206\,679\,894\,179\,894\,2$ $3.589\,955\,357\,142\,857\,142\,9$	
NOT CONVERGED	$\ell = 0.012\,704\,475\,963\,893\,30$ $0.294\,868\,000\,440\,917\,107\,58$ $-2.663\,168\,540\,564\,373\,897\,7$ $10.701\,467\,702\,821\,869\,489$ $-25.124\,736\,000\,881\,834\,215$ $38.020\,414\,462\,081\,128\,748$ $-38.540\,361\,000\,881\,834\,215$ $26.310\,842\,702\,821\,869\,489$ $-11.884\,150\,683\,421\,516\,755$ $3.884\,823\,357\,583\,774\,250\,4$
NOT CONVERGED	NOT CONVERGED

Numerical experiments

The purpose of the experiment is to verify accuracy and stability properties of the stabilised Adams-type methods constructed above. We also display results of the classic implicit Adams methods of corresponding orders, which have longer stability intervals than their classical explicit counterparts. In all our experiments we use constant step size and reference solutions computed by *Wolfram Mathematica's* `NDSolve`. The starting points were taken from this reference solution. For each method we perform a series of constant-step integrations with decreasing step size τ and calculate the maximum norm of the error at the endpoint. Missing points on the convergence diagrams mean that the error is too large due to instability of the method for the particular value of τ.

HIRES. This is a classical mildly stiff test system of dimension 8 describing a chemical reaction (see [1, chapter IV.10, formula (10.4)]). All equations except the 6th and 7th are linear. The interval of integration is [0, 40]. Figure 3, *a*, shows the performance of the 6-step stabilised methods of orders 1–6 and the implicit method of order 6. We see that the results agree well with common sense: more accurate methods have shorter stability intervals. Then we compare methods of order 5 and display the results at fig. 3, *b*, where we took k from 9 to 15 in order to get larger stability intervals than the implicit method have. There is clear evidence that methods with larger k have larger error constants. We do not show the results of the damped first order method, since the difference compared to the simple non-damped methods is negligible for this test problem.

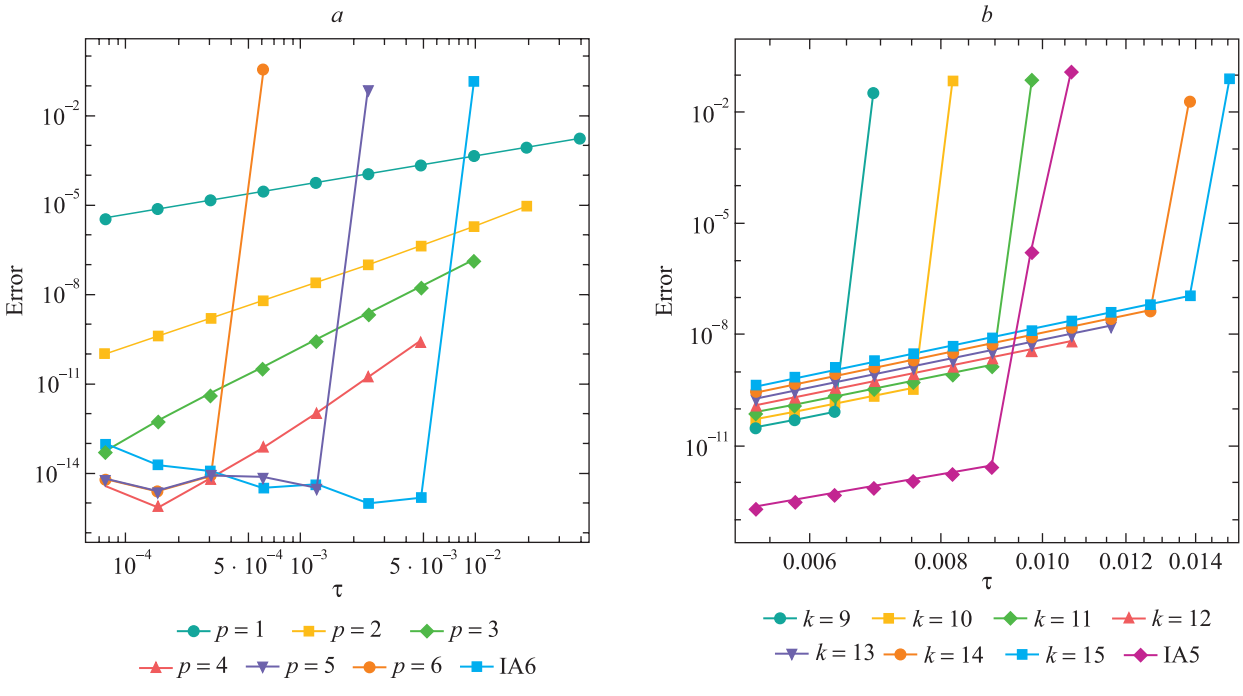

Fig. 3. Numerical experiment with HIRES problem.
Six-step stabilised methods and implicit Adams method of order 6 (*a*).
Stabilised methods of order 5 and implicit Adams method of the same order (*b*)

Burgers' equation. The second problem is taken from [5]. Consider a method of lines discretisation of the one-dimensional non-linear boundary value problem

$$\begin{aligned} u_t + \left(\frac{u^2}{2}\right)_x &= \mu u_{xx},\ x \in [0, 1],\ t \in [0, 0.25], \\ u(x, 0) &= 1.5x(1-x)^2, \\ u(0, t) &= u(1, t) = 0. \end{aligned} \tag{18}$$

The spatial derivatives are approximated by standard central finite differences, the discretisation step is $\Delta x = \frac{1}{501}$, so the dimension of the resulting ordinary differential equation is 500. The Jacobi matrix of this problem is not symmetric and complex eigenvalues occur for sufficiently small values of μ. We took $\mu = 0.005$ for which this is not the case at the starting point, but apparently non-real eigenvalues do emerge during integration.

The first experiment is similar to the one from the previous problem. The results are presented on fig. 4, *a*: we compare six-step methods of different orders. We see that the first order method with damping allows for taking longer time steps than the non-damped one. This indicates that the solution generates non-real eigenvalues of the Jacobi matrix. Hence, it is unlikely to benefit from using stabilized methods with large k and p, for which we do not have damping yet.

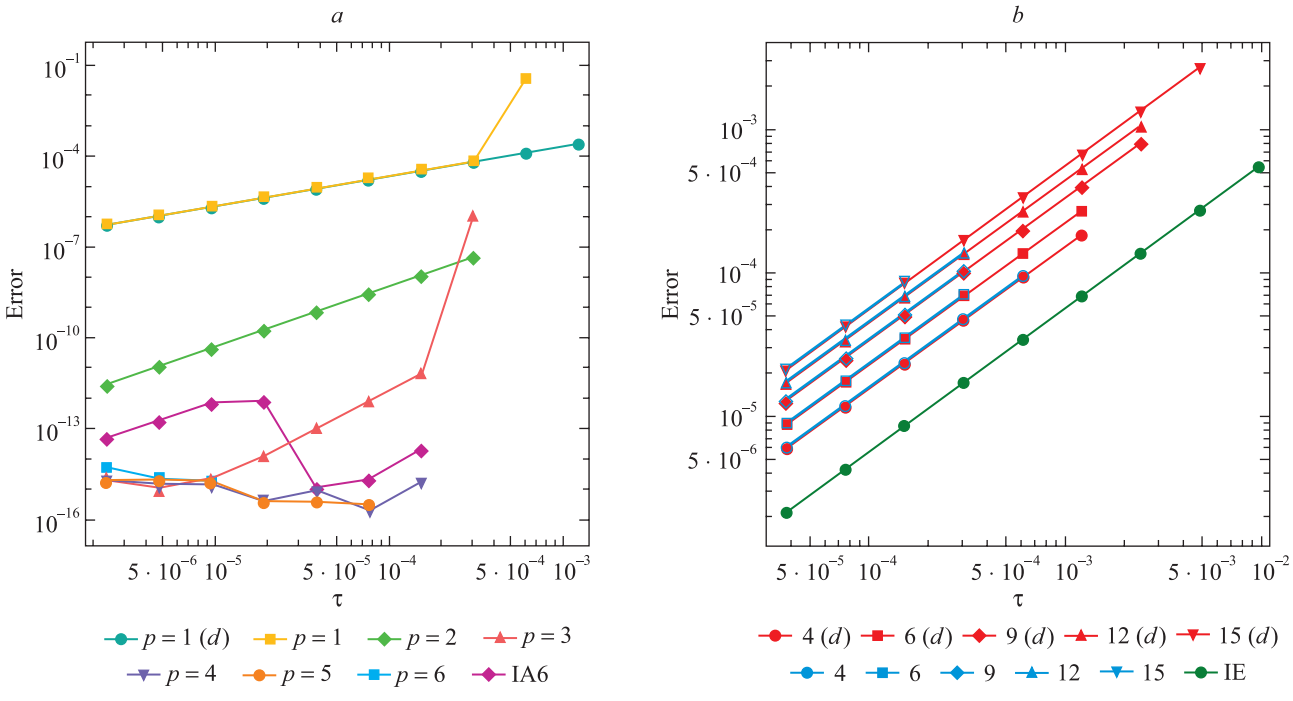

Fig. 4. Numerical experiment with Burgers' equation (18).
Six-step stabilised methods and implicit Adams method of order 6 (*a*).
First order stabilised methods with and without damping
for $\varepsilon = 0.25$, $k = 4, 6, 9, 12, 15$, and the implicit Euler method (*b*)

Indeed, our experiments showed that these methods cannot take larger steps than implicit Adams-type methods of the same order, even if their stability interval is longer. Hence, on fig. 4, *b*, we compare only stabilised explicit methods of order one with the implicit Euler method. This experiment shows that damping is crucial for the general performance of a stabilised method. Another obvious conclusion is that for this particular problem the explicit methods are less accurate than the implicit one.

Conclusion

In this work we presented explicit multi-step methods of Adams type, which possess extended stability intervals. Simple formulae for the first order methods and their error constants are derived. We also applied damping to the first order methods, derived a general scheme for construction of stabilised k-step methods of any order $p < k$, and calculated coefficients for such methods numerically. It was shown that the error constant of the stabilised method grows as the number of steps increases, but this growth is quite slow. Our numerical experiments asserted the theoretical properties of accuracy and stability of the constructed methods and exhibited the importance of damping transformation for the methods.

In our opinion, at present the stabilised Adams-type methods are mostly of theoretical interest, but it cannot be ruled out that they could be useful in practice and become a basis for a competitive solver for mildly stiff problems. From a practical perspective the methods are attractive due to their low cost (just one evaluation of f per step for any k and p) and simplicity of implementation. The weak point is that long stability intervals require a large number of steps, which will entail memory issues, difficulties with the starting values and so on.

Appendix A

Mathematica code for computing the stabilised method's parameters

```
ClearAll[a, b, beta, oc, mu]
k = 5; o = 3;
mu[betas_List] := With[{k = Length@betas}
      , Evaluate[(#^k - #^(k - 1))/(#^Range[0, k - 1]).betas] &
      ];
param = {a[-1] -> 0, a[k] -> 0
      , a[k - 1] -> Sum[b[j]^2, {j, 0, k - 1}]
      , a[j_] :> Sum[b[l] b[l + k - j - 1], {l, 0, j}]
      , beta[k - 1] -> a[k - 1] + a[k - 2]
      , beta[j_] :> a[j - 1] + a[j]
      };
bs = b /@ Range[0, k - 1];
betas = beta /@ Range[0, k - 1];
oc[1] = Total@betas - 1;
oc[p_] := Simplify[betas.Range[1 - k, 0]^(p - 1)] - 1/p;
cons = Thread[(oc /@ Range[o] //. param) == 0];
sol = NMinimize[Prepend[cons, bs.bs], bs
      , Method -> Automatic
      , WorkingPrecision -> 50
      , AccuracyGoal -> 25
      , PrecisionGoal -> 25
      , MaxIterations -> 1000
      ];
betaopt = (betas //. param) /. sol[[2]];
rescond = (oc /@ Range[o]) /. Thread[betas -> betaopt];
<|"k" -> k, "order" -> o, "betas" -> betaopt, "orderres" -> rescond,
"len" -> -mu[betaopt][-1]|>
```

References

1. Hairer E, Wanner G. *Solving ordinary differential equations II: stiff and differential-algebraic problems.* Berlin: Springer; 1996. 614 p. (Springer series in computational mathematics; volume 14). DOI: 10.1007/978-3-642-05221-7.

2. Lebedev VI. How to solve stiff systems of differential equations by explicit methods. In: Marchuk GI, editor. *Numerical methods and applications.* Boca Raton: CRC Press; 1994. p. 45–80.

3. Sommeijer BP, Shampine LF, Verwer JG. RKC: an explicit solver for parabolic PDEs. *Journal of Computational and Applied Mathematics.* 1998;88(2):315–326. DOI: 10.1016/S0377-0427(97)00219-7.

4. Abdulle A, Medovikov AA. Second order Chebyshev methods based on orthogonal polynomials. *Numerische Mathematik.* 2001;90(1):1–18. DOI: 10.1007/s002110100292.

5. Abdulle A. Fourth order Chebyshev methods with recurrence relation. *SIAM Journal on Scientific Computing.* 2002;23(6): 2041–2054. DOI: 10.1137/S1064827500379549.

6. Jeltsch R, Nevanlinna O. Stability of explicit time discretizations for solving initial value problems. *Numerische Mathematik.* 1981;37(1):61–91. DOI: 10.1007/BF01396187.

7. Jeltsch R, Nevanlinna O. Stability and accuracy of time discretizations for initial value problems. *Numerische Mathematik.* 1982;40(2):245–296. DOI: 10.1007/BF01400542.

8. Daubechies I. *Ten lectures on wavelets.* Philadelphia: Society for Industrial and Applied Mathematics; 1992. 369 p. (CBMS-NSF regional conference series in applied mathematics).

9. Hairer E, Nørsett SP, Wanner G. *Solving ordinary differential equations I: nonstiff problems.* 2nd edition. Berlin: Springer; 1993. 528 p. (Springer series in computational mathematics; volume 8). DOI: 10.1007/978-3-540-78862-1.

10. Xu Y, Zhao JJ. Estimation of longest stability interval for a kind of explicit linear multistep methods. *Discrete Dynamics in Nature and Society.* 2010;2010:1–18. DOI: 10.1155/2010/912691.

Received 22.02.2021 / revised 08.06.2021 / accepted 08.06.2021.

УДК 539.32:536.2

ВЛИЯНИЕ ПРОТЯЖЕННЫХ ИСТОЧНИКОВ ТЕПЛА НА РАСПРЕДЕЛЕНИЕ ТЕМПЕРАТУРЫ В ПРОФИЛИРОВАННЫХ ПОЛЯРНО-ОРТОТРОПНЫХ КОЛЬЦЕВЫХ ПЛАСТИНАХ С ТЕПЛОИЗОЛИРОВАННЫМИ ОСНОВАНИЯМИ

***В. В. КОРОЛЕВИЧ*[1)], *Д. Г. МЕДВЕДЕВ*[2)]**

[1)]*Международный центр современного образования, ул. Штепанска, 61, 110 00, г. Прага 1, Чехия*
[2)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Приводится решение стационарной задачи теплопроводности для профилированных полярно-ортотропных кольцевых пластин с теплоизолированными основаниями от N протяженных источников тепла на их внешних границах. Распределение температур в таких пластинах является неосесимметричным. Решение стационарной задачи теплопроводности для анизотропных кольцевых пластин произвольного профиля записывается через решение соответствующего интегрального уравнения Вольтерры 2-го рода. Представлена формула расчета температур в анизотропных кольцевых пластинах произвольного профиля. Получено точное решение стационарной задачи теплопроводности для полярно-ортотропной кольцевой пластины степенного профиля. Показано, что в такой анизотропной пластине распределение температуры от N протяженных источников тепла на ее внешней границе имеет более сложный характер, чем в случае распределения температуры от N точечных источников тепла на ее внешней границе.

Ключевые слова: полярно-ортотропная кольцевая пластина; температура; стационарное уравнение теплопроводности; интегральное уравнение Вольтерры 2-го рода; пластина степенного профиля.

Образец цитирования:
Королевич ВВ, Медведев ДГ. Влияние протяженных источников тепла на распределение температуры в профилированных полярно-ортотропных кольцевых пластинах с теплоизолированными основаниями. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:99–104.
https://doi.org/10.33581/2520-6508-2021-2-99-104

For citation:
Karalevich UV, Medvedev DG. Influence of extended heat sources on the temperature distribution in profiled polar-orthotropic annular plates with heat-insulated bases. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:99–104. Russian.
https://doi.org/10.33581/2520-6508-2021-2-99-104

Авторы:
Владимир Васильевич Королевич – преподаватель.
Дмитрий Георгиевич Медведев – доктор педагогических наук, кандидат физико-математических наук, доцент; первый проректор.

Authors:
Uladzimir V. Karalevich, lecturer.
v.korolevich@mail.ru
Dmitrij G. Medvedev, doctor of science (pedagogics), PhD (physics and mathematics), docent; first vice-rector.
medvedev@bsu.by

Журнал Белорусского государственного университета. Математика. Информатика. 2021;2:99–104
Journal of the Belarusian State University. Mathematics and Informatics. 2021;2:99–104

INFLUENCE OF EXTENDED HEAT SOURCES ON THE TEMPERATURE DISTRIBUTION IN PROFILED POLAR-ORTHOTROPIC ANNULAR PLATES WITH HEAT-INSULATED BASES

U. V. KARALEVICH[a], *D. G. MEDVEDEV*[b]

[a]*International Center of Modern Education, 61 Štěpánská Street, Prague 1, PSČ 110 00, Czech*
[b]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: U. V. Karalevich (v.korolevich@mail.ru)

The solution of the stationary heat conduction problem for profiled polar-orthotropic annular plates with heat-insulated bases from *N* extended heat sources at their external border is presented. The temperature distribution in such plates will be non-axisymmetric. The solution of the stationary heat conduction problem for anisotropic annular plates of an random profile is resolved through the solution of the corresponding Volterra integral equation of the second kind. The formula of a temperature calculations in anisotropic annular plates of an random profile is given. The exact solution of stationary heat conduction problem for polar-orthotropic annular plate of an exponential profile is recorded. The temperature distribution in such anisotropic plate from *N* extended heat sources at its outer border is more complex than in the case of temperature distribution from *N* point heat sources at their external border.

Keywords: polar-orthotropic annular plate; temperature; stationary equation of heat conduction; Volterra integral equation of the second kind; plate of an exponential profile.

Введение

В данной статье исследуется неосесимметричное распределение температуры $T(r,\theta)$ в полярно-ортотропных кольцевых пластинах переменной толщины $h(r)$ с теплоизолированными основаниями, когда на внутреннем контуре ($r = r_0$) пластины поддерживается постоянная температура T_1^*, а на внешнем контуре ($r = R$) приложены N источников тепла с температурой T_2^* каждый.

В отличие от работы [1] в настоящей статье учитывается *дискретность* расположения *протяженных источников тепла* на внешнем контуре анизотропной кольцевой пластины. Полученные результаты имеют практическую значимость при проектировании и расчете на прочность профилированных кольцевых пластин аппаратов пищевой и химической промышленности, изготавливаемых из современных композитных материалов. Возникающие в них термоупругие напряжения могут существенно влиять на напряженно-деформированное состояние кольцевых пластин аппарата. Расчетная схема, учитывающая протяженность конечного числа источников тепла на внешней границе профилированных анизотропных кольцевых пластин, позволяет реально оценить вклад термоупругих напряжений в общую картину распределения напряжений в данных кольцевых пластинах аппаратов пищевых и химических производств.

Постановка задачи и основные уравнения

В работе исследуется влияние протяженности источников тепла на внешней границе на распределение температуры в анизотропной кольцевой пластине переменной толщины, основания которой при $z = \pm\dfrac{h}{2}$ теплоизолированы. Теплообмен нагретой тонкой кольцевой пластины с внешней средой через боковую цилиндрическую поверхность пренебрежимо мал, и его можно не учитывать в расчетах. Предполагается, что температура в тонкой кольцевой пластине не меняется по толщине. Внутренних источников тепла в ней не имеется, а тепловое поле является плоским и неосесимметричным. Теплофизические характеристики материала пластины будем полагать постоянными и не зависящими от температуры.

Конечно, идеальных точечных источников тепла в природе не существует. Все они имеют какую-то протяженность. В работе [2] нами получено распределение температуры на внешнем контуре $T^{\text{внеш}}$, если N протяженных источников тепла с температурой T_2^* каждый приложены на равноотстоящих одинаковых дугах длиной l с центральным углом φ ($l = \varphi R$):

$$T^{\text{внеш}}(R,\theta,\varphi) = NT_2^*\left(1 + 2\sum_{n=1}^{\infty}\frac{\sin n\varphi}{n\varphi}\cos Nn\theta\right).$$

Ниже центральным углом φ будет определяться протяженность источника тепла на внешней границе кольцевой пластины.

Количество N источников тепла не может быть произвольным, а ограничивается температурой плавления $T_{\text{плавл}}$ материала пластины: $N_{\max} = \left[\dfrac{T_{\text{плавл}}}{T_2^*} \right]$, где квадратные скобки означают целую часть дробного выражения.

В цилиндрической системе координат r, θ, z уравнение стационарной теплопроводности для полярно-ортотропной кольцевой пластины переменной толщины $h(r)$ с теплоизолированными основаниями имеет вид [3]

$$\lambda_r \frac{\partial^2 T}{\partial r^2} + \left(\frac{h'(r)}{h(r)} + 1 \right) \lambda_r \frac{\partial T}{\partial r} + \frac{1}{r^2} \lambda_\theta \frac{\partial^2 T}{\partial \theta^2} = 0, \tag{1}$$

где λ_r и λ_θ – радиальный и тангенциальный коэффициенты теплопроводности материала пластины, не зависящие от температуры $T(r, \theta, \varphi)$.

Разложим функцию $T(r, \theta, \varphi)$ в тригонометрический ряд Фурье:

$$T(r, \theta, \varphi) = T_0(r) + \sum_{n=1}^{\infty} T_{Nn}^{(1)}(r, \varphi) \cos nN\theta + \sum_{n=1}^{\infty} T_{Nn}^{(2)}(r, \varphi) \sin nN\theta. \tag{2}$$

Первое слагаемое $T_0(r)$ в разложении (2) описывает осесимметричное распределение температуры в пластине. Слагаемые, содержащие $\cos nN\theta$, соответствуют симметричным составляющим функции $T(r, \theta, \varphi)$ относительно плоскости $\theta = 0$, а слагаемые, содержащие $\sin nN\theta$, – обратно симметричным.

Подстановка разложения (2) в уравнение (1) приводит к системе обыкновенных дифференциальных уравнений для компонент $T_0(r)$, $T_{Nn}^{(i)}(r, \varphi)$ ($i = 1, 2$):

$$\begin{cases} (n = 0) & \dfrac{d^2 T_0}{dr^2} + \left(\dfrac{h'(r)}{h(r)} + \dfrac{1}{r} \right) \dfrac{dT_0}{dr} = 0, \qquad (3) \\ (n \geq 1) & \dfrac{d^2 T_{Nn}^{(i)}}{dr^2} + \left(\dfrac{h'(r)}{h(r)} + \dfrac{1}{r} \right) \dfrac{dT_{Nn}^{(i)}}{dr} - \dfrac{\lambda_\theta}{\lambda_r} \dfrac{(Nn)^2}{r^2} T_{Nn}^{(i)}(r, \varphi) = 0. \qquad (4) \end{cases}$$

Получим теперь граничные условия для функции температуры $T(r, \theta, \varphi)$.

$$\begin{cases} T(r_0, \theta, \varphi) = T_0(r_0) + \sum\limits_{n=1}^{\infty} T_n^{(1)}(r_0, \varphi) \cos Nn\theta + \sum\limits_{n=1}^{\infty} T_n^{(2)}(r_0, \varphi) \sin Nn\theta = \\ = T^{\text{внутр}}(r_0, \theta) = T_1^* + \sum\limits_{n=1}^{\infty} 0 \cdot \cos Nn\theta + \sum\limits_{n=1}^{\infty} 0 \cdot \sin Nn\theta, \\ T(R, \theta, \varphi) = T_0(R) + \sum\limits_{n=1}^{\infty} T_n^{(1)}(R, \varphi) \cos Nn\theta + \sum\limits_{n=1}^{\infty} T_n^{(2)}(R, \varphi) \sin Nn\theta = \\ = T^{\text{внеш}}(R, \theta, \varphi) = NT_2^* + \sum\limits_{n=1}^{\infty} 2NT_2^* \dfrac{\sin n\varphi}{n\varphi} \cos Nn\theta + \sum\limits_{n=1}^{\infty} 0 \cdot \sin Nn\theta. \end{cases}$$

Из сравнения коэффициентов тригонометрических рядов при одинаковых гармониках левых и правых частей приведенных выражений следуют граничные условия

$$(n = 0) \quad \begin{cases} T_0(r_0) = T_1^*, \\ T_0(R) = NT_2^*, \end{cases} \tag{5}$$

$$(n \geq 1) \quad \begin{cases} T_{Nn}^{(1)}(r_0, \varphi) = 0, \\ T_{Nn}^{(1)}(R, \varphi) = 2NT_2^* \dfrac{\sin n\varphi}{n\varphi}, \end{cases} \tag{6}$$

$$(n \geq 1) \quad \begin{cases} T_{Nn}^{(2)}(r_0, \varphi) = 0, \\ T_{Nn}^{(2)}(R, \varphi) = 0. \end{cases} \tag{7}$$

При нулевых граничных условиях (7) однородное дифференциальное уравнение (4) имеет тривиальное решение, следовательно, функция $T_{Nn}^{(2)}(r, \varphi)$ равна нулю, т. е. обратно симметричная составляющая в разложении (2) для функции $T(r, \theta, \varphi)$ отсутствует.

Обыкновенное дифференциальное уравнение (3) легко интегрируется, и его решение при заданных граничных условиях (5) имеет вид

$$T_0(r) = \left(1 - \frac{\int_{r_0}^{r} \frac{ds}{sh(s)}}{\int_{r_0}^{R} \frac{ds}{sh(s)}}\right) T_1^* + \frac{\int_{r_0}^{r} \frac{ds}{sh(s)}}{\int_{r_0}^{R} \frac{ds}{sh(s)}} NT_2^*.$$

Решение неосесимметричной задачи стационарной теплопроводности методом линейных интегральных уравнений Вольтерры 2-го рода

Для профилированной анизотропной кольцевой пластины общее решение обыкновенного дифференциального уравнения (4) системы выразим через решение соответствующего ему интегрального уравнения Вольтерры 2-го рода:

$$T_{Nn}^{(1)}(r, \varphi) = \int_{r_0}^{r} (r - s) \eta_{Nn}^{(1)}(s, \varphi) ds + \dot{T}_{Nn}^{(1)}(r_0, \varphi)(r - r_0) + T_{Nn}^{(1)}(r_0, \varphi). \tag{8}$$

Здесь разрешающая функция $\eta_{Nn}^{(1)}(r, \varphi)$ удовлетворяет интегральному уравнению Вольтерры 2-го рода:

$$\eta_{Nn}^{(1)}(r, \varphi) = \lambda \int_{r_0}^{r} K_{Nn}(r, s) \eta_{Nn}^{(1)}(s, \varphi) ds + f_{Nn}^{(1)}(r, \varphi), \tag{9}$$

где $\lambda = -1$ есть числовой параметр; $K_{Nn}(r, s) = \frac{1}{r} + \frac{h'(r)}{h(r)} - \frac{\lambda_\theta}{\lambda_r} \frac{(Nn)^2}{r^2}(r - s)$ – ядро интегрального уравнения; $f_{Nn}^{(1)}(r, \varphi) = \frac{\partial K_{Nn}(r, s)}{\partial s} T_{Nn}^{(1)}(r_0, \varphi) - K_{Nn}(r, r_0) \dot{T}_{Nn}^{(1)}(r_0, \varphi)$ – свободный член интегрального уравнения.

Общее решение интегрального уравнения Вольтерры 2-го рода (9) записывается с помощью *резольвенты* $R_{Nn}(r, s; \lambda)$ в виде [4]

$$\eta_{Nn}^{(1)}(r, \varphi) = \lambda \int_{r_0}^{r} R_{Nn}(r, s; \lambda) f_{Nn}^{(1)}(s, \varphi) ds + f_{Nn}^{(1)}(r, \varphi),$$

где функция $R_{Nn}(r, s; \lambda)$ определяется функциональным рядом

$$R_{Nn}(r, s; \lambda) = \sum_{m=0}^{\infty} \lambda^m K_{Nn, m+1}(r, s),$$

который для непрерывных итерированных ядер $K_{Nn, m}(r, s)$ сходится абсолютно и равномерно.

Используя граничные условия (6), найдем постоянные $T_{Nn}^{(1)}(r_0, \varphi)$, $\dot{T}_{Nn}^{(1)}(r_0, \varphi)$:

$$\begin{cases} T_{Nn}^{(1)}(r_0, \varphi) = 0, \\ T_{Nn}^{(1)}(R, \varphi) = \int_{r_0}^{R} (R - s) \eta_{Nn}^{(1)}(s, \varphi) ds + \dot{T}_{Nn}^{(1)}(r_0, \varphi)(R - r_0) = 2NT_2^* \frac{\sin n\varphi}{n\varphi}. \end{cases}$$

Отсюда

$$\begin{cases} T_{Nn}^{(1)}(r_0, \varphi) = 0, \\ \dot{T}_{Nn}^{(1)}(r_0, \varphi) = \frac{1}{R - r_0} \left(-\int_{r_0}^{R} (R - s) \eta_{Nn}^{(1)}(s, \varphi) ds + 2NT_2^* \frac{\sin n\varphi}{n\varphi} \right). \end{cases}$$

Окончательное выражение для компонент $T_{Nn}^{(1)}(r, \varphi)$ имеет вид

$$T_{Nn}^{(1)}(r, \varphi) = \int_{r_0}^{r} (r - s)\eta_{Nn}^{(1)}(s, \varphi)ds - \frac{r - r_0}{R - r_0}\int_{r_0}^{R} (R - s)\eta_{Nn}^{(1)}(s, \varphi)ds + 2NT_2^* \frac{r - r_0}{R - r_0}\frac{\sin n\varphi}{n\varphi}. \quad (10)$$

Подставляя формулы (8) и (10) для компонент $T_0(r)$, $T_{Nn}^{(1)}(r, \varphi)$ в разложение (2), получаем в общем случае распределение температуры $T(r, \theta, \varphi)$ в профилированной анизотропной кольцевой пластине с теплоизолированными основаниями от протяженных источников тепла на ее внешней границе:

$$T(r, \theta, \varphi) = \left(1 - \frac{\int_{r_0}^{r} \frac{ds}{sh(s)}}{\int_{r_0}^{R} \frac{ds}{sh(s)}}\right) T_1^* + \left\{ \frac{\int_{r_0}^{r} \frac{ds}{sh(s)}}{\int_{r_0}^{R} \frac{ds}{sh(s)}} + 2\left(\frac{r - r_0}{R - r_0}\right)\sum_{n=1}^{\infty} \frac{\sin n\varphi}{n\varphi}\cos nN\theta + \right.$$

$$\left. + \sum_{n=1}^{\infty}\left[\int_{r_0}^{r} (r - s)\eta_{Nn}^{(1)}(s, \varphi)ds - \frac{(r - r_0)}{(R - r_0)}\int_{r_0}^{R} (R - s)\eta_{Nn}^{(1)}(s, \varphi)ds\right]\cos nN\theta \right\} NT_2^*. \quad (11)$$

Для полярно-ортотропной кольцевой пластины, толщина которой меняется по степенному закону $h(r) = h_0\left(\frac{r_0}{r}\right)^{\alpha}$, где $\alpha \in R \setminus \{0\}$ и h_0 – толщина пластины на внутреннем контуре при $r = r_0$, система обыкновенных дифференциальных уравнений (3), (4) имеет точное решение, удовлетворяющее граничным условиям (5), (6):

$$\begin{cases} T_0(r) = \left(\frac{1 - \left(\frac{r}{R}\right)^{\alpha}}{1 - \left(\frac{r_0}{R}\right)^{\alpha}}\right) T_1^* + \left(\frac{\left(\frac{r}{R}\right)^{\alpha} - \left(\frac{r_0}{R}\right)^{\alpha}}{1 - \left(\frac{r_0}{R}\right)^{\alpha}}\right) NT_2^*, \\ T_{Nn}^{(1)}(r, \varphi) = -2NT_2^*\left[\frac{\delta^{k_2(n)}}{\left(\delta^{k_1(n)} - \delta^{k_2(n)}\right)}\left(\frac{r}{R}\right)^{k_1(n)} - \frac{\delta^{k_1(n)}}{\left(\delta^{k_1(n)} - \delta^{k_2(n)}\right)}\left(\frac{r}{R}\right)^{k_2(n)}\right]\frac{\sin n\varphi}{n\varphi}. \end{cases} \quad (12)$$

Здесь $\delta = \frac{r_0}{R}$, $k_{1,2}(n) = \frac{\alpha}{2} \pm \sqrt{\left(\frac{\alpha}{2}\right)^2 + \frac{\lambda_\theta}{\lambda_r}(Nn)^2}$ – корни характеристического уравнения.

Введем безразмерную координату $x = \frac{r}{R}$ и подставим решения (12) в разложение (2). В результате имеем неосесимметричное распределение температуры в полярно-ортотропной кольцевой пластине степенного профиля с теплоизолированными основаниями от N протяженных источников тепла на ее внешней границе:

$$T(x, \theta, \varphi) = \left(\frac{1 - x^{\alpha}}{1 - \delta^{\alpha}}\right) T_1^* +$$

$$+ \left\{ \left(\frac{x^{\alpha} - \delta^{\alpha}}{1 - \delta^{\alpha}}\right) - 2\sum_{n=1}^{\infty}\left[\frac{\delta^{k_2(n)}}{\left(\delta^{k_1(n)} - \delta^{k_2(n)}\right)}x^{k_1(n)} - \frac{\delta^{k_1(n)}}{\left(\delta^{k_1(n)} - \delta^{k_2(n)}\right)}x^{k_2(n)}\right]\frac{\sin n\varphi}{n\varphi}\cos nN\theta \right\} NT_2^*. \quad (13)$$

Совершая предельный переход $\varphi \to 0$ $\left(\lim\limits_{\varphi \to 0}\frac{\sin n\varphi}{n\varphi} = 1\right)$ в формулах (11), (13), получаем неосесимметричное распределение температуры в анизотропных кольцевых пластинах переменной толщины от N точечных источников тепла на ее внешней границе.

Выводы

В профилированных анизотропных кольцевых пластинах распределение температуры от N протяженных источников тепла на внешней границе имеет более сложный характер, чем распределение температуры от N точечных источников тепла на внешней границе. Поскольку формулы (11), (13) содержат ряды, в которые входит тригонометрическая функция $\sin n\varphi$, то эти ряды будут знакопеременными. Более того, ввиду быстрого стремления к нулю осциллирующей функции $\frac{\sin n\varphi}{n\varphi}$ при увеличении n в формулах (11), (13) при практических расчетах можно ограничиться только несколькими первыми членами рядов.

Библиографические ссылки

1. Королевич ВВ, Медведев ДГ. Решение неосесимметричной стационарной задачи теплопроводности для полярно-ортотропной кольцевой пластины переменной толщины с теплоизолированными основаниями. *Журнал Белорусского государственного университета. Математика. Информатика*. 2018;1:77–87.

2. Королевич ВВ, Медведев ДГ. Влияние протяженности источников тепла на внешней границе на распределение температуры в профилированных полярно-ортотропных кольцевых пластинах с учетом теплообмена с окружающей средой. *Журнал Белорусского государственного университета. Математика. Информатика*. 2020;3:86–91. DOI: 10.33581/2520-6508-2020-3-86-91.

3. Уздалев АИ. *Некоторые задачи термоупругости анизотропного тела.* Саратов: Издательство Саратовского университета; 1967. 167 с.

4. Краснов МЛ, Киселев АИ, Макаренко ГИ. *Интегральные уравнения: задачи и примеры с подробными решениями*. Москва: КомКнига; 2007. 192 с.

References

1. Karalevich UV, Medvedev DG. The solution of the nonaxisymmetric stationary problem of heat conduction for the polar-orthotropic annular plate of variable thickness with thermal insulated bases. *Journal of the Belarusian State University. Mathematics and Informatics.* 2018;1:77–87. Russian.

2. Karalevich UV, Medvedev DG. The influence of the length of heat sources on the external border on the temperature distribution in profiled polar-orthotropic ring plates taking into account there heat exchange with the external environment. *Journal of the Belarusian State University. Mathematics and Informatics.* 2020;3:86–91. Russian. DOI: 10.33581/2520-6508-2020-3-86-91.

3. Uzdalev AI. *Nekotorye zadachi termouprugosti anizotropnogo tela* [Some problems of thermoelasticity of an anisotropic body]. Saratov: Izdatel'stvo Saratovskogo universiteta; 1967. 167 p. Russian.

4. Krasnov ML, Kiselev AI, Makarenko GI. *Integral'nye uravneniya: zadachi i primery s podrobnymi resheniyami* [Integral equations: problems and examples with detailed solutions]. Moscow: KomKniga; 2007. 192 p. Russian.

Получена 04.05.2021 / исправлена 02.06.2021 / принята 02.06.2021.
Received 04.05.2021 / revised 02.06.2021 / accepted 02.06.2021.

УДК 531.31;539.43;539.411.5

АНАЛИТИЧЕСКАЯ МОДЕЛЬ ДВИЖЕНИЯ СКИПА С УЧЕТОМ НАЛИЧИЯ ГОЛОВНОГО И УРАВНОВЕШИВАЮЩЕГО КАНАТОВ

М. А. ЖУРАВКОВ[1)]***, В. П. САВЧУК***[1)]***, М. А. НИКОЛАЙЧИК***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Приведена процедура построения аналитической модели, описывающей динамику движения шахтного скипа с учетом наличия головного и уравновешивающего канатов и криволинейности проводников. Выведены формулы для сил, действующих на скип со стороны проводников. Показано, что частоты собственных колебаний скипа зависят от вертикального ускорения и пройденного пути при подъеме сосуда. Построен график (диаграмма) вертикальной скорости скипа, соблюдение которого не вызывает появления значимых вертикальных колебаний скипа из-за упругости канатов. Разработан алгоритм нахождения главного вектора и главного момента системы сил, действующих на скип, по показаниям трех акселерометров, регистрирующих горизонтальные ускорения сосуда при его движении.

Ключевые слова: скип; проводники; динамика шахтного скипа; вертикальные колебания; главный вектор системы сил; главный момент системы сил; горизонтальные ускорения.

ANALYTICAL MODEL OF SKIP MOTION TAKING INTO ACCOUNT INFLUENCE OF HEAD AND BALANCING ROPES

M. A. ZHURAVKOV[a]***, V. P. SAVCHUK***[a]***, M. A. NIKOLAITCHIK***[a]

[a]*Belarusian State University, 4 Niezalieznasci Avenue, Minsk 220030, Belarus*
Corresponding author: M. A. Nikolaitchik (nikolaitchik.m@gmail.com)

The article describes an analytical model of mine skip dynamics taking into account the presence of the head and balancing ropes and the existing curvilinearity of the guides. Expressions for the forces acting on the skip from the side of the guides have been constructed. It is shown, that the frequencies of natural vibrations of skip depend on the vertical acceleration and the distance traveled during its lifting. A graph (diagram) of skips vertical speed which observance does not lead to the appearance of skips vertical vibrations due to elasticity of the ropes is developed. An algorithm for finding the forces principal vector and the forces principal moment acting on the skip based on the reading of three accelerometers recording horizontal accelerations of skip during its movement is presented.

Keywords: skip; guides; mine skip dynamics; vertical vibrations; forces principal vector; forces principal moment; horizontal accelerations.

Образец цитирования:
Журавков МА, Савчук ВП, Николайчик МА. Аналитическая модель движения скипа с учетом наличия головного и уравновешивающего канатов. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:105–113 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-105-113

For citation:
Zhuravkov MA, Savchuk VP, Nikolaitchik MA. Analytical model of skip motion taking into account influence of head and balancing ropes. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:105–113.
https://doi.org/10.33581/2520-6508-2021-2-105-113

Авторы:
Михаил Анатольевич Журавков – доктор физико-математических наук, профессор; заведующий кафедрой теоретической и прикладной механики механико-математического факультета.
Владимир Петрович Савчук – кандидат физико-математических наук; доцент кафедры теоретической и прикладной механики механико-математического факультета.
Михаил Александрович Николайчик – заведующий научно-исследовательской лабораторией прикладной механики механико-математического факультета.

Authors:
Michael A. Zhuravkov, doctor of science (physics and mathematics), full professor; head of the department of theoretical and applied mechanics, faculty of mechanics and mathematics.
zhuravkov@bsu.by
Vladimir P. Savchuk, PhD (physics and mathematics); associate professor at the department of theoretical and applied mechanics, faculty of mechanics and mathematics.
Mikhail A. Nikolaitchik, head of the laboratory of applied mechanics, faculty of mechanics and mathematics.
nikolaitchik.m@gmail.com

Introduction

A simplified scheme of shaft lifting rig adopted in the article consists of a friction pulley driven by the lifting machine, two skips moving along the guides and the main and the counterweight rope. Four spring-loaded rollers are installed at the end of the skip frame and they copy profiles of the guides during its movement. Guides are welded box-section beams with local deviations from the vertical in two horizontal directions. Ascent of the loaded skip and the descent of the empty one is carried out by the head rope slung over the friction pulley. Unloading the skip in its upper position and loading the empty skip in its lower position are performed simultaneously after which the pulley changes direction of rotation and the working cycle repeats.

System lifting vessel – reinforcement dynamics investigation during the lifting vessel movement and this system dynamic behaviour diagnostics is an urgent applied problem [1–3]. At the same time this problem is very complex in terms of correct mechanics and mathematical models construction and an adequate model analysis performing.

The research objectives presented in this article were:

• construction of the skip motion analytical model with an acceptable degree of accuracy. That model must allow to determine the force effect on the skip from the side of rigid curved guides or according to their known profile, or according to the readings of accelerometers installed on the skip;

• determination of the skip natural horizontal vibrations frequencies;

• the hoisting machine torque change graph obtaining, which excludes skip vertical vibrations occurrence in an elastic rope.

Skip motion vector equations

The skip motion as a mechanical system can be represented as a rigid body complex motion (*cxyz* coordinate system), consisting of translational motion of $O_1X_1Y_1Z_1$ coordinate system with a given speed $v(t)$ along OZ axis of the stationary system $OXYZ$ ($v(t)$ is the vertical speed of the skip mass center) and of five independent movements of system *cxyz*: two translational movements along O_1X_1 and O_1Y_1 axes and three rotations around axes CX', CY', CZ' of Koenig coordinate system (directions of the corresponding axes $OXYZ$, $O_1X_1Y_1Z_1$ and $CX'Y'Z'$ coincide). The resulting complex movement which occurs relative to the given bulk motion of the system $O_1X_1Y_1Z_1$ is obtained as a result of the superposition of these five motions. Each of these motions and consequently the resulting one arises due to external horizontal influences on skip from the side of the guides and vertical influences at points M_0 and M_5 of the main and balancing ropes suspension (fig. 1).

The skip – guide contact node at contact points M_i, $i = \overline{1, 4}$, is schematically represented as consisting of three independent springs in contact with three guide surfaces through rollers that roll along the conductor during the skip motions (fig. 2).

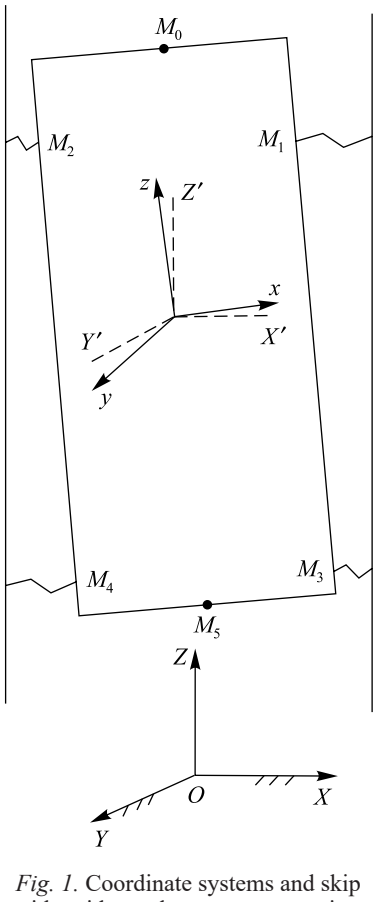

Fig. 1. Coordinate systems and skip with guides and ropes contact points

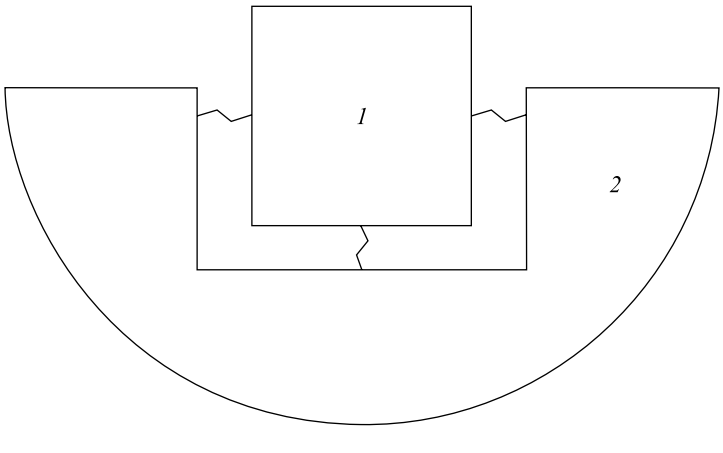

Fig. 2. Skip with guides contact node scheme: *1* – guide; *2* – skip

One can use the following system of equations to describe the skip with mass M relative motion dynamic under the action of the forces $\bar{F}_i$ applied at points M_i, $i=\overline{0,5}$:

$$M\bar{W}_c=\sum_{i=1}^{4}\bar{F}_i, \tag{1}$$

$$\frac{d\bar{K}_c'}{dt}=\sum_{i=0}^{5}\overline{CM_i'}\times\bar{F}_i. \tag{2}$$

Equations (1), (2) are vector notation of the skip mass center horizontal motion relative to $O_1X_1Y_1Z_1$ coordinate system and the change in angular momentum $\bar{K}'$ relative to the Koenig coordinate system [4].

In equations (1), (2)

$$\bar{F}_i=\left(F_{ix},\,F_{iy},\,0\right),\ i=\overline{1,4},\ \bar{F}_0=\left(0,\,0,\,F_{0Z}\right),\ \bar{F}_5=\left(0,\,0,\,F_{5Z}\right),$$

$$\overline{CM_i'}=\overline{CM_i}+\left(\varphi_x,\,\varphi_y,\,\varphi_z\right)\times\overline{CM_i},\ i=\overline{0,5},$$

$$\overline{CM_0}=\left(0,\,0,\,z_0\right),\ \overline{CM_5}=\left(0,\,0,\,z_5\right),\ \overline{CM_1}=\left(x_1,\,0,\,z_1\right),$$

$$\overline{CM_2}=\left(-x_1,\,0,\,z_1\right),\ \overline{CM_3}=\left(x_1,\,0,\,z_3\right),\ \overline{CM_4}=\left(-x_1,\,0,\,z_3\right),$$

where φ_x, φ_y, φ_z are the angles between the corresponding axes of coordinate systems $cxyz$ and $CX'Y'Z'$. Dashes indicate that the vectors belong to the coordinate system $CX'Y'Z'$.

Considering skip as an absolutely rigid body symmetric with respect to planes cxz, cyz, and angles φ_x, φ_y, φ_z small one can replace the moments of inertia of skip I_x', I_y', I_z' relative to the axes of system $CX'Y'Z'$ in expression for $\bar{K}_c'$ with the moments of inertia I_x, I_y, I_z relative to the axes of the coordinate system $cxyz$: $\bar{K}_c'=\left(I_x\dot{\varphi}_x,\,I_y\dot{\varphi}_y,\,I_z\dot{\varphi}_z\right)$.

Forces acting on a skip during its motion

The spring-loaded rollers of the contact node copy the guide cylindrical surfaces from its three sides during skip motion. The equations of these surfaces in the $OXYZ$ fixed coordinate system can be written for the first guide as

$$X=h_x+f_{1x}(z),\ Y=h_y+f_{1y}(z),\ Y=-h_y+f_{1y}(z)$$

and for the second guide as

$$X=-h_x+f_{2x}(z),\ Y=h_y+f_{2y}(z),\ Y=-h_y+f_{2y}(z).$$

Functions $f_{1x}(z)$, $f_{1y}(z)$, $f_{2x}(z)$, $f_{2y}(z)$ give the deviations algebraic value of the guides points from the corresponding vertical planes $X=\pm h_x$, $Y=\pm h_y$.

The force values F_{ix}, F_{iy}, $i=\overline{1,4}$, are equal to the corresponding springs compression (tension) to the multiplied by the stiffness coefficient c which is considered the same for all springs since the guides act on the skip through the springs. Small displacements $\Delta\rho_i$, $i=\overline{1,4}$, of the skip points M_i relative to the coordinate system $O_1X_1Y_1Z_1$ occur to the mass center displacement by a vector $\left(X_c,\,Y_c,\,0\right)$ and three rotations around the mass center by angles φ_x, φ_y, φ_z. Thus

$$\Delta\bar{\rho}_1=\left(X_c,\,Y_c,\,0\right)+\left(\varphi_x,\,\varphi_y,\,\varphi_z\right)\times\left(x_1,\,0,\,z_1\right)=\left(X_c+z_1\varphi_y,\,Y_c+x_1\varphi_z-z_1\varphi_x,\,-x_1\varphi_y\right).$$

Similarly, we have

$$\Delta\bar{\rho}_2=\left(X_c+z_1\varphi_y,\,Y_c-x_1\varphi_z-z_1\varphi_x,\,x_1\varphi_y\right),$$

$$\Delta\bar{\rho}_3=\left(X_c+z_3\varphi_y,\,Y_c+x_1\varphi_z-z_3\varphi_x,\,-x_1\varphi_y\right),$$

$$\Delta\bar{\rho}_4=\left(X_c+z_3\varphi_y,\,Y_c-x_1\varphi_z-z_3\varphi_x,\,x_1\varphi_y\right).$$

Taking into account the conductors deviations, we obtain

$$F_{1x}=c\left(f_{1x}(s+h)-X_c-z_1\varphi_y\right),\ F_{2x}=c\left(f_{2x}(s+h)-X_c-z_1\varphi_y\right),$$

$$F_{3x} = c\left(f_{1x}(s) - X_c - z_3\varphi_y\right),\ F_{4x} = c\left(f_{2x}(s) - X_c - z_3\varphi_y\right),$$
$$F_{1y} = 2c\left(f_{1y}(s+h) - Y_c - x_1\varphi_z + z_1\varphi_x\right),\ F_{2y} = 2c\left(f_{2y}(s+h) - Y_c + x_1\varphi_z + z_1\varphi_x\right),$$
$$F_{3y} = 2c\left(f_{1y}(s) - Y_c - x_1\varphi_z + z_3\varphi_x\right),\ F_{4y} = 2c\left(f_{2y}(s) - Y_c + x_1\varphi_z + z_3\varphi_x\right).$$

Here $s(t) = \int_0^t v(\tau)d\tau$ is the distance passed by the skip during by the time t after the its movement start, $h = z_1 - z_3$ is the distance between points M_1 and M_3.

A part of the balancing rope from the suspension point to the loop in the sump, which length is approximately $s(t)$, moves translationally with a speed $v(t)$ together with the skip (fig. 3). Thus

$$F_{0Z} = \left(M + \rho_b s(t)\right)\left(\dot{v}(t) + g\right),\ F_{5Z} = -\rho_b s(t)\left(\dot{v}(t) + g\right),$$

where ρ_b is the rope density and g is the gravity acceleration.

Skip motion scalar equations and their solution

We project the vector equation (1) on the horizontal axes of $OXYZ$ coordinate system and equation (2) on the Koenig coordinate system axis to obtain a system of skip motion scalar equations

$$M\ddot{X}_c = c\left(F_x - 4X_c - 2z_{13}\varphi_y\right), \tag{3}$$

$$M\ddot{Y}_c = 2c\left(F_y - 4Y_c + 2z_{13}\varphi_x\right), \tag{4}$$

$$I_x\ddot{\varphi}_x = 2c\left(\Phi_x + 2z_{13}Y_c - 2\left(z_1^2 + z_3^2\right)\varphi_x\right) - F_{05}\varphi_x, \tag{5}$$

$$I_y\ddot{\varphi}_y = c\left(\Phi_y - 2z_{13}X_c - 2\left(z_1^2 + z_3^2\right)\varphi_y\right) - F_{05}\varphi_y, \tag{6}$$

$$I_z\ddot{\varphi}_z = 2cx_1\left(\Phi_z - 4x_1\varphi_z\right). \tag{7}$$

$$F_x = f_{1x}(s+h) + f_{2x}(s+h) + f_{1x}(s) + f_{2x}(s),$$
$$F_y = f_{1y}(s+h) + f_{2y}(s+h) + f_{1y}(s) + f_{2y}(s),$$
$$\Phi_x = -z_1\left(f_{1y}(s+h) + f_{2y}(s+h)\right) - z_3\left(f_{1y}(s) + f_{2y}(s)\right),$$
$$\Phi_y = z_1\left(f_{1x}(s+h) + f_{2x}(s+h)\right) + z_3\left(f_{1x}(s) + f_{2x}(s)\right),$$
$$\Phi_z = f_{1y}(s+h) + f_{1y}(s) - f_{2y}(s+h) - f_{2y}(s),$$
$$F_{05} = z_0F_{0Z} + z_5F_{5Z},\ z_{13} = z_1 + z_3.$$

Note that equations (3)–(7) approximate such as deriving them, discarded terms of higher order terms to the values X_c, Y_c, φ_x, φ_y, φ_z. In addition note that equations (5), (6) have variable coefficients due to the function $F_{05}(t)$ presence. Analysis of this function form shows that its change during the skip movement is much slower than changing the desired system functions (3)–(7).

An approximate solution of equations with slowly varying coefficients is usually found by the asymptotic averaging method [5]. Following this method in our case an approximate solution can be obtained by considering the variable coefficients «frozen», i. e. constant when performing the solution. The «frozen» coefficients «unfreeze» and their dependence on time is restored after the system analytical solution obtaining.

The system (3)–(7) solution with «frozen» coefficients is easily found by using the integral Laplace transform [6] and has the form

$$\varphi_z(t) = \sqrt{\frac{c}{2I_z}} \cdot \int_0^t \Phi_z(\tau)\sin\omega_1(t-\tau)d\tau,\ \omega_1 = 2x_1\sqrt{\frac{2c}{I_z}},$$

$$X_c(t)=\int_0^t A_1(t-\tau)F_x(\tau)d\tau-\int_0^t A_2(t-\tau)\Phi_y(\tau)d\tau,$$

$$\varphi_y(t)=\int_0^t A_3(t-\tau)\Phi_y(\tau)\,d\tau-\int_0^t A_2(t-\tau)F_x(\tau)\,d\tau,$$

$$Y_c(t)=\int_0^t B_1(t-\tau)F_y(\tau)d\tau+\int_0^t B_2(t-\tau)\Phi_x(\tau)d\tau, \quad (8)$$

$$\varphi_x(t)=\int_0^t B_3(t-\tau)\Phi_x(\tau)d\tau+\int_0^t B_2(t-\tau)F_y(\tau)d\tau;$$

$$A_i(t)=a_{i1}\sin\omega_2 t+a_{i2}\sin\omega_3 t,\ i=\overline{1,3},$$

$$a_{11}=\frac{c\left(2c\left(z_1^2+z_3^2\right)-I_y\omega_2^2+F_{05}\right)}{a_5},$$

$$a_{12}=\frac{c\left(2c\left(z_1^2+z_3^2\right)-I_y\omega_3^2+F_{05}\right)}{a_6},$$

$$a_5=\omega_2\left(a_2-2a_1\omega_2^2\right),\ a_6=\omega_3\left(a_2-2a_1\omega_3^2\right),$$

$$a_1=MI_y,\ a_2=2Mc\left(z_1^2+z_3^2\right)+4cI_y+MF_{05},$$

$$a_3=4c^2\left(z_1-z_3\right)^2+4cF_{05},\ \omega_2=\sqrt{\frac{a_2-a_4}{2a_1}},\ \omega_3=\sqrt{\frac{a_2+a_4}{2a_1}},$$

$$a_4=\sqrt{a_2^2-4a_1a_3},\ a_{21}=\frac{2c^2z_{13}}{a_5},\ a_{22}=\frac{2c^2z_{13}}{a_6},$$

$$a_{31}=c\frac{4c-M\omega_2^2}{a_5},\ a_{32}=c\frac{4c-M\omega_3^2}{a_6};$$

$$B_i(t)=b_{i1}\sin\omega_4 t+b_{i2}\sin\omega_5 t,\ i=\overline{1,3},$$

$$b_{11}=2c\frac{4c\left(z_1^2+z_3^2\right)-I_x\omega_4^2+F_{05}}{b_5},\ b_{12}=2c\frac{4c\left(z_1^2+z_3^2\right)-I_x\omega_5^2+F_{05}}{b_6},$$

$$b_1=MI_x,\ b_2=4Mc\left(z_1^2+z_3^2\right)+8cI_x+MF_{05},\ b_3=16c^2\left(z_1-z_3\right)^2+8cF_{05},$$

$$b_4=\sqrt{b_2^2-4b_1b_3},\ \omega_4=\sqrt{\frac{b_2-b_4}{2b_1}},\ \omega_5=\sqrt{\frac{b_2+b_4}{2b_1}},$$

$$b_5=\omega_4\left(b_2-2b_1\omega_4^2\right),\ b_6=\omega_5\left(b_2-2b_1\omega_5^2\right),$$

$$b_{21}=\frac{8c^2z_{13}}{b_5},\ b_{22}=\frac{8c^2z_{13}}{b_6},\ b_{31}=2c\frac{8c-M\omega_4^2}{b_5},\ b_{32}=2c\frac{8c-M\omega_5^2}{b_6}.$$

The formulas obtained allow us to find the force effect to the skip at points M_i, $i = \overline{0, 5}$, according to the guides known profiles, the acceleration $\dot{v}(t)$ and the path $s(t)$. Note that only ω_1 of skip natural vibrations five frequencies ω_i, $i = \overline{1, 5}$, has a constant value. The remaining frequencies depend on the functions $\dot{v}(t)$, $s(t)$, which in turn are determined by the moment $M_{cr}(t)$ applied on the mine hoisting machine drum.

The relationship between speed v and torque M_{cr}

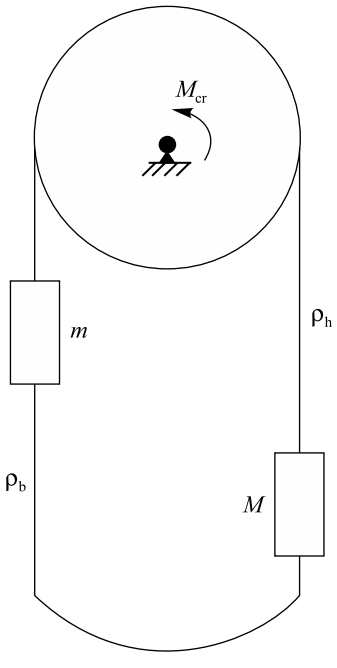

Fig. 3. Mine hoist scheme

An approximate relationship between the quantities $v(s)$ and $M_{cr}(s)$ considering the head and the balancing ropes inextensible can be obtained from the theorem on the change in kinetic energy [4] according to which the kinetic energy differential of the mine hoisting mechanism is equal to the elementary work sum of the gravity forces and the torque $M_{cr}(s)$ (see fig. 3):

$$\frac{1}{2} M_{eff}\, d\left(v^2(s)\right) = \left(\frac{M_{cr}(s)}{r} + \left(m - M + l\rho_b - l\rho_h\right) g - 2g\left(\rho_b - \rho_h\right) s \right) ds. \quad (9)$$

Here $M_{eff} = m + M + \left(\rho_b + \rho_h\right) l + \frac{I}{r^2}$, m is the empty skip mass; ρ_b, ρ_h are the balancing and the head ropes linear densities, respectively; I, r are the moment of inertia and drum radius; l is the the ropes length; $0 \le s(t) \le l$.

From equality (9) it follows that the loaded skip will begin to rise under the condition

$$M_{cr}(0) > rg\left(M - m + l\rho_h - l\rho_b\right).$$

After dividing equality (9) by dt we obtain an equation

$$M_{eff}\ddot{s}(t) = \frac{M_{cr}(t)}{r} + \left(m - M + l\rho_b - l\rho_h\right) g - 2g\left(\rho_b - \rho_h\right) s(t). \quad (10)$$

Assuming that the loaded skip lifting takes place under the constant torque M_{cr} action from a rest state at $s = 0$ and with a stop at the upper point $s = l$ then from equation (9) we obtain that this is possible only at

$$M_{cr} = \left(M - m\right) gr$$

and

$$v^2(s) = \frac{2gs\left(\rho_b - \rho_h\right)(l - s)}{M_{eff}}. \quad (11)$$

From (11) it follows that condition $\rho_b > \rho_h$ must be satisfied and the maximum speed value is reached when $s = \frac{l}{2}$:

$$v_{max} = l\sqrt{\frac{g\left(\rho_b - \rho_h\right)}{2M_{eff}}}.$$

Solving equation (10) at $M_{cr} = \left(M - m\right) gr$ taking into account $s(0) = 0$, $\dot{s}(0) = 0$ we obtain

$$s(t) = \frac{l}{2}\left(1 - \cos\left(\sqrt{\frac{2g\left(\rho_b - \rho_h\right)}{M_{eff}}} \cdot t\right)\right). \quad (12)$$

From (12) we receive the skip motion time t_0 to a complete stop:

$$t_0 = \pi\sqrt{\frac{M_{eff}}{2g\left(\rho_b - \rho_h\right)}}.$$

In general case $M_{cr}(t)$ is obtained from equation (10) according to the given motion law $s(t)$ and is processed by the mine hoisting machine digital control system. Therefore, it is necessary to ensure the second derivative $\ddot{s}(t)$ continuity on the interval $[0, t_0]$ and its equality to zero at the segment extreme points to obtain a continuous change in $M_{cr}(t)$.

A smooth function $s(t)$ constructing example

Let us construct function $s(t)$ with five sections of its second derivative linear variation:

$$\ddot{s}(t) = 4v_{\max}\begin{cases} \dfrac{t}{t_1^2},\ 0 \le t \le \dfrac{t_1}{2}, \\ \dfrac{t_1 - t}{t_1^2},\ \dfrac{t_1}{2} \le t \le t_1, \\ 0,\ t_1 \le t \le t_0 - t_2, \\ \dfrac{t_0 - t_2 - t}{t_2^2},\ t_0 - t_2 \le t \le t_0 - \dfrac{t_2}{2}, \\ \dfrac{t - t_0}{t_2^2},\ t_0 - \dfrac{t_2}{2} \le t \le t_0. \end{cases}$$

Here $v_{\max}$ is the maximum motion speed; $t_1 = \dfrac{2v_{\max}}{w_1}$, $t_2 = \dfrac{2v_{\max}}{w_2}$ are the time of acceleration and deceleration, t_0 is the motion time; w_1, w_2 are the largest acceleration modules during acceleration and deceleration.

Integration gives the speed $v(t)$ and path $s(t)$:

$$v(t) = v_{\max}\begin{cases} \dfrac{2t^2}{t_1^2},\ 0 \le t \le \dfrac{t_1}{2}, \\ 1 - \dfrac{2(t_1 - t)^2}{t_1^2},\ \dfrac{t_1}{2} \le t \le t_1, \\ 1,\ t_1 \le t \le t_0 - t_2, \\ 1 - \dfrac{2(t_0 - t_2 - t)^2}{t_2^2},\ t_0 - t_2 \le t \le t_0 - \dfrac{t_2}{2}, \\ \dfrac{2(t - t_0)^2}{t_2^2},\ t_0 - \dfrac{t_2}{2} \le t \le t_0. \end{cases}$$

$$s(t) = v_{\max}\begin{cases} \dfrac{2t^3}{3t_1^2},\ 0 \le t \le \dfrac{t_1}{2}, \\ t + \dfrac{2(t_1 - t)^3}{3t_1^2} - \dfrac{t_1}{2},\ \dfrac{t_1}{2} \le t \le t_1, \\ t - \dfrac{t_1}{2},\ t_1 \le t \le t_0 - t_2, \\ t + \dfrac{2(t_0 - t_2 - t)^3}{3t_2^2} - \dfrac{t_1}{2},\ t_0 - t_2 \le t \le t_0 - \dfrac{t_2}{2}, \\ \dfrac{2(t - t_0)^3}{3t_2^2} - \dfrac{t_2}{2} + t_0 - \dfrac{t_1}{2},\ t_0 - \dfrac{t_2}{2} \le t \le t_0. \end{cases}$$

If l is the distance of the skip lift then $s(t_0) = l$ and the motion time is $t_0 = \dfrac{l}{v_{\max}} + \dfrac{t_1 + t_2}{2}$.

The principal vector and the principal moment of the forces acting on the skip

If twice differentiate function (8) we obtain the principal vector $\bar{F}_c$ and the principal moment $\bar{M}_c$ of the forces acting on a skip in its relative motion:

$$\bar{F}_c = \left(M\ddot{X}_c,\, M\ddot{Y}_c,\, 0\right),\ \ \bar{M}_c = \left(I_x\ddot{\varphi}_x,\, I_y\ddot{\varphi}_y,\, I_z\ddot{\varphi}_z\right). \tag{13}$$

In this case the guides profiles and all skip mechanical characteristics must be known. Also the components of the mass center acceleration $\bar{W}_c = \left(\ddot{X}_c,\, \ddot{Y}_c,\, 0\right)$ and angular acceleration $\bar{\varepsilon} = \left(\ddot{\varphi}_x,\, \ddot{\varphi}_y,\, \ddot{\varphi}_z\right)$ in equalities (13) can be expressed in terms of the accelerations horizontal components $\bar{W}_i$, $i = \overline{6,\, 8}$, of any three skip points M_i, $i = \overline{6,\, 8}$, that do not lie on one straight line and accelerations $\bar{W}_i$ can be obtained from the readings of the accelerometers in points M_i. Indeed if we take as a pole for example a point M_6 then taking into account the first quantities order we obtain [4]

$$\bar{W}_c = \bar{W}_6 + \bar{\varepsilon} \times \overline{M_6C},$$

$$\bar{\varepsilon} \times \overline{M_6M_7} = \bar{W}_7 - \bar{W}_6, \tag{14}$$

$$\bar{\varepsilon} \times \overline{M_6M_8} = \bar{W}_8 - \bar{W}_6. \tag{15}$$

Vector equalities (14), (15) give the following four scalar equalities when projected onto horizontal axes:

$$\begin{aligned}
&\ddot{\varphi}_y\left(z_7 - z_6\right) - \ddot{\varphi}_z\left(y_7 - y_6\right) = W_{7x} - W_{6x},\\
&\ddot{\varphi}_z\left(x_7 - x_6\right) - \ddot{\varphi}_x\left(z_7 - z_6\right) = W_{7y} - W_{6y},\\
&\ddot{\varphi}_y\left(z_8 - z_6\right) - \ddot{\varphi}_z\left(y_8 - y_6\right) = W_{8x} - W_{6x},\\
&\ddot{\varphi}_z\left(x_8 - x_6\right) - \ddot{\varphi}_x\left(z_8 - z_6\right) = W_{8y} - W_{6y}.
\end{aligned} \tag{16}$$

One can compose a uniquely solvable system for finding $\ddot{\varphi}_x$, $\ddot{\varphi}_y$, $\ddot{\varphi}_z$ from these equalities. In particular, from the first three equations (16) we obtain

$$\ddot{\varphi}_x = \frac{\Delta_x}{\Delta},\ \ddot{\varphi}_y = \frac{\Delta_y}{\Delta},\ \ddot{\varphi}_z = \frac{\Delta_z}{\Delta},$$

$$\Delta = \left(z_7 - z_6\right)\left(\left(y_7 - y_6\right)\left(z_8 - z_6\right) - \left(y_8 - y_6\right)\left(z_7 - z_6\right)\right),$$

$$\Delta_x = -\left(W_{7x} - W_{6x}\right)\left(z_8 - z_6\right)\left(x_7 - x_6\right) + \left(W_{8x} - W_{6x}\right)\left(z_7 - z_6\right) \times$$

$$\times \left(x_7 - x_6\right) + \left(W_{7y} - W_{6y}\right)\left(\left(z_7 - z_6\right)\left(y_8 - y_6\right) - \left(z_8 - z_6\right)\left(y_7 - y_6\right)\right),$$

$$\Delta_y = \left(z_7 - z_6\right)\left(\left(W_{8x} - W_{6x}\right)\left(y_7 - y_6\right) - \left(W_{7x} - W_{6x}\right)\left(y_8 - y_6\right)\right),$$

$$\Delta_z = \left(z_7 - z_6\right)\left(\left(W_{8x} - W_{6x}\right)\left(z_7 - z_6\right) - \left(W_{7x} - W_{6x}\right)\left(z_8 - z_6\right)\right).$$

Coordinates of points M_i, $i = \overline{6,\, 8}$, must not vanish the determinant Δ.

Conclusion

An approximate analytical model for the mine skip motion has been developed. It is important that this model takes into account the head and balance ropes influence as well as the curvature of the guides.

Expressions that determine the force interaction of skip with guides during the lifting vessel motion are obtained. The expressions obtained make it possible to determine the principal vector and the principal moment of the forces acting on the skip using data coming from three accelerometers installed on the skip fixing the horizontal accelerations values.

An analysis of the skip natural vibrations frequencies has been carried out. Expressions that determine the dependence of the natural vibrations frequencies on the vertical acceleration and the path covered by the skip are given. A diagram of the speed, that does not provoke the vertical vibrations occurrence of skip on the elastic rope is proposed.

Библиографические ссылки

1. Самуся ВИ, Ильина ИС, Ильина СС. Компьютерное моделирование и исследование динамики систем «сосуд – армировка» в стволах с нарушенной геометрией. *Вестник ПНИПУ. Геология. Нефтегазовое и горное дело.* 2016;15(20):277–285. DOI: 10.15593/2224-9923/2016.20.8.

2. Ильин СР, Дубинин МВ. Спектральный и деформационный анализ систем «сосуд – армировка» вертикальных стволов. *Геотехническая механика.* 2015;122:164–193.

3. Трифанов ГД, Зверев ВЮ, Вагин ЕО, Архипов ЕВ. Оценка влияния кинематических параметров подъемных установок на динамические нагрузки в канатах. *Горный информационно-аналитический бюллетень.* 2017;7:103–110.

4. Савчук ВП, Медведев ДГ, Вярвильская ОН. *Теоретическая механика.* Минск: БГУ; 2016. 231 с.

5. Моисеев НН. *Асимптотические методы нелинейной механики.* Москва: Наука; 1969. 380 с.

6. Снеддон ИН. *Преобразования Фурье.* Матвеев АН, переводчик. Москва: Издательство иностранной литературы; 1955. 667 с.

References

1. Samusia VI, Iliina IS, Iliina SS. Computer modeling and investigation of dynamics of system «vessel – reinforcement» in shafts with broken geometry. *Bulletin of PNRPU. Geology Oil & Gas Engineering & Mining.* 2016;15(20):277–285. DOI: 10.15593/2224-9923/2016.20.8. Russian.

2. Ilyin SR, Dubinin MV. Spectral and deformation analysis of systems «vessels – reinforcement» of vertical shaft. *Geo-Technical Mechanics.* 2015;122:164–193. Russian.

3. Trifanov GD, Zverev VYu, Vagin EO, Archipov EV. Estimation of influence exerted by kinematic parameters of hoisting machines on the dynamics loads in cables. *Gornyi informatsionno-analiticheskii byulleten'.* 2017;7:103–110. Russian.

4. Savchuk VP, Medvedev DG, Vyarvil'skaya ON. *Teoreticheskaya mekhanika.* Minsk: Belarusian State University; 2016. 231 p. Russian.

5. Moiseev NN. *Asimptoticheskie metody nelineinoi mekhaniki.* Moscow: Science; 1969. 380 p. Russian.

6. Sneddon IN. *Fourier transforms.* Moscow: McGraw-Hill; 1951. 542 p.

Russian edition: Sneddon IN. *Preobrazovaniya Fur'e.* Matveev AN, translator. Moscow: Izdatel'stvo inostrannoi literatury; 1955. 667 p.

Received 17.03.2021 / revised 23.06.2021 / accepted 23.06.2021.

Теоретические основы информатики

Theoretical foundations of computer science

УДК 004:932

ИДЕНТИФИКАЦИЯ ОБЪЕКТОВ ЗЕМНОЙ ПОВЕРХНОСТИ НА ОСНОВЕ АНСАМБЛЕЙ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Е. Е. МАРУШКО[1)]***, А. А. ДУДКИН***[1)]***, С. ЧЕН***[2)]

[1)]*Объединенный институт проблем информатики НАН Беларуси, ул. Сурганова, 6, 220012, г. Минск, Беларусь*
[2)]*Сианьский институт оптики и точной механики Китайской академии наук, Шэньси, 710119, г. Сиань, Китай*

В работе предлагается методика идентификации объектов на изображениях поверхности Земли, основанная на сочетании методов машинного обучения. В качестве исходных моделей рассматриваются различные варианты многослойных сверточных нейронных сетей и машин опорных векторов. Предлагается также гибридная сверточная нейронная сеть, которая комбинирует признаки, выделенные нейронной сетью и экспертами. Оптимальные значения гиперпараметров моделей вычисляются методами сеточного поиска с использованием *k*-кратной перекрестной проверки. Показана возможность повышения точности идентификации на основе ансамблей указанных моделей нейронных сетей. Эффективность предложенного подхода демонстрируется на примере изображений, полученных радаром с синтезированной апертурой.

Образец цитирования:
Марушко ЕЕ, Дудкин АА, Чен С. Идентификация объектов земной поверхности на основе ансамблей сверточных нейронных сетей. *Журнал Белорусского государственного университета. Математика. Информатика.* 2021;2:114–123 (на англ.).
https://doi.org/10.33581/2520-6508-2021-2-114-123

For citation:
Marushko EE, Doudkin AA, Zheng X. Identification of Earth's surface objects using ensembles of convolutional neural networks. *Journal of the Belarusian State University. Mathematics and Informatics.* 2021;2:114–123.
https://doi.org/10.33581/2520-6508-2021-2-114-123

Авторы:
Евгений Евгеньевич Марушко – научный сотрудник лаборатории идентификации систем.
Александр Арсентьевич Дудкин – доктор технических наук, профессор; заведующий лабораторией идентификации систем.
Сиантао Чен – кандидат наук (обработка сигналов и информации); доцент базовой лаборатории технологий спектральной обработки изображений.

Authors:
Evgenii E. Marushko, researcher at the laboratory of identification systems.
marushkoee@gmail.com
Alexander A. Doudkin, doctor of science (engineering), full professor; head of the laboratory of identification systems.
doudkin@newman.bas-net.by
Xiangtao Zheng, PhD (signal and information processing); associate professor at the key laboratory of spectral imaging technology.
xiangtaoz@gmail.com

Ключевые слова: сверточная нейронная сеть; машина опорных векторов; ансамбль нейронных сетей; изображение поверхности Земли; дистанционное зондирование; радар с синтезированной апертурой.

Благодарность. Работа частично поддержана Белорусским фондом фундаментальных исследований и национальным фондом естественных наук Китая (проект № Ф20-017).

IDENTIFICATION OF EARTH'S SURFACE OBJECTS USING ENSEMBLES OF CONVOLUTIONAL NEURAL NETWORKS

E. E. MARUSHKO[a]***, A. A. DOUDKIN***[a]***, X. ZHENG***[b]

[a]*United Institute of Informatics Problems, National Academy of Sciences of Belarus, 6 Surhanava Street, Minsk 220012, Belarus*
[b]*Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Shaanxi, Xi'an 710119, China*
Corresponding author: E. E. Marushko (marushkoee@gmail.com)

The paper proposes an identification technique of objects on the Earth's surface images based on combination of machine learning methods. Different variants of multi-layer convolutional neural networks and support vector machines are considered as original models. A hybrid convolutional neural network that combines features extracted by the neural network and experts is proposed. Optimal values of hyperparameters of the models are calculated by grid search methods using *k*-fold cross-validation. The possibility of improving the accuracy of identification based on the ensembles of these models is shown. Effectiveness of the proposed technique is demonstrated by the example of images obtained by synthetic aperture radar.

Keywords: convolutional neural network; support vector machine; neural network ensemble; Earth's surface image; remote sensing; identification; synthetic aperture radar.

Acknowledgements. This work was partially supported by the Belarusian Republican Foundation for Fundamental Research and the National Foundation of Natural Sciences of China (project No. F20-017).

Introduction

Remote sensing of the Earth is an observation of the Earth's surface by ground, aviation and space imaging means. Wavelengths received by the imaging equipment ranges from visible optical to radio waves. A multi-channel equipment of the passive and active types is used, that registers an electromagnetic radiation. Passive sensing methods use the natural reflected or secondary thermal radiation of Earth's objects due to solar activity. Active methods use stimulated emission of the objects, initiated by an artificial source. Remote sensing data is characterised by a large degree of dependence on the transparency of the atmosphere.

Digital data are represented as a two-dimensional image for each spectral range in the form of a matrix (two-dimensional array) of numbers $I(i, j)$ of the intensity of radiation received by a sensor from elements of the Earth's surface, which correspond to pixels of the image, where (i, j) are coordinates of the pixels. If the image is obtained in several ranges of the electromagnetic spectrum, it is represented by a three-dimensional matrix consisting of the numbers $I(i, j, k)$, where k is the number of the spectral channel. Thus, the information obtained during remote sensing is data with spatial relationships between the features $I(i, j)$.

Artificial neural networks (NNs) are successfully applied for solving image processing problems including object identification. Researches in the field of increasing the efficiency of identification based on NN theory are carried out in the following two main directions.

1. Development of the unique most appropriate multi-layer hybrid NN model for object identification on images which combining some popular NN models to solve the realistic identification process efficiently. This model is constructed from at least two different types of NNs. The first part of the architecture is aimed to image pre-processing and feature extraction, the second – directly to object detection. There are known different NN combinations for this goal: multi-layer perceptron, convolutional neural network (CNN), self-organising map, long short-term memory, NN realisation of principle component analysis, several support vector machines (SVMs), recurrent NN, etc. [1–4].

2. Development of ensembles of neural networks (ENN). They are sets of NNs making decision by averaging the results of individual NNs improving the identification quality [3; 5; 6].

In recent years, deep NNs have been most successfully used for processing on images obtained by Earth remote sensing [7]. Our work combines the above-mentioned approaches by using an ensemble of hybrid CNNs and SVMs to solve an image identification problem: to discern the nature of image components (objects). The main contribution of the paper is to distinguish the objects of the known class from the objects of alien classes (one-class classification). Effectiveness of the proposed technique is demonstrated on the task to identify objects of two classes on images obtained by synthetic aperture radar.

Methods and algorithms

The architecture of CNNs was proposed by LeCun [8] and it is aimed at effective image recognition. The NN architecture got its name because of convolution operations, where each image fragment is multiplied by the convolution matrix (kernel) element by element, and the result is summed up and written to the same position [9]. CNN is usually an alternation of convolutional layers, pooling layers, dense layers and an output layer. Additionally, a dropout layers can be used.

The dense (fully connected) layer connects each neuron with all neurons at the previous level. Each connection has its own weight. In the convolution layer (in contrast to the dense layer), a neuron is connected only with a limited number of neurons of the previous level. The convolutional layer is similar to the use of the convolution operation, which uses only a weight matrix of a small size (convolution kernel). The pooling layer performs dimension reduction. This can be done in various ways, but the method of selecting the maximum element is often used – the previous layer output is divided into cells, the maximum value is selected in each cell that is transferred to the next layer. The dropout layer is a matrix of coefficients and can be used together with all mentioned layers for a weight regularisation. The regularisation (dropout) consists in changing the NN structure: each neuron turns off with a certain probability at a stage of a training using stochastic gradient descent. The training is performed on a thinned NN, and a gradient step is made for the remaining weights. The output layer performs a class of identified objects.

The accuracy of identifying objects can be improved using ENNs [10–12]. It is necessary to realise the variability of NNs in the ensemble. The following approaches or their combinations can be applied for this purpose:

- using different parts of the training set;
- random initialisation of NN weights;
- variation of NN architectures in the ensemble (adding hidden layers, adding or deleting neurons of the hidden layer).

The output value of the ensemble is formed as a weighted sum of the outputs of the individual NNs. This approach is illustrated in fig. 1. For the case with one output neuron, the result is calculated by the equation

$$y = \sum_{i=1}^{n} y_i w_i,$$

where n is the number of the NNs; y_i is the output of the i NN; w_i is the weight of the i NN, which is calculated by the formula

$$w_i = \frac{A_i}{\sum_{j=1}^{n} A_j},$$

where A_i is a chosen measure of error calculated for the i NN; n is the number of NNs.

SVM is one of the most popular supervised learning methods proposed by Vapnik [13]. It creates a hyperplane or a set of hyperplanes in a multi-dimensional space that can be used to solve problems of classification, regression and other close problems. If the data is linearly inseparable, a non-linear kernel is applied, which allows to map the source data to a space of higher dimension, where an optimal separating hyperplane can exist.

The method is recommended for processing a small set of features. Therefore, it can be chosen as the main method for forming a model from manually selected features of objects in the image. Thus, CNNs that receive input data directly in the form of images and SVMs that make decisions on selected features of objects can be combined into an ensemble (fig. 2). This scheme can be modified by submitting additional features, formed without using images, directly to the input of the SVM classifiers. CNN can be modified similarly. The network can be divided into several branches for data processing (fig. 3).

One branch performs automatic feature extraction on the image using standard CNN layers, the weighting coefficients of which are determined by gradient methods during training. The other branch may include a set of predetermined processing procedures to form an additional set of features for each input image. Also, the sets of external features can be submitted to the hybrid model. This model involves two stages of training.

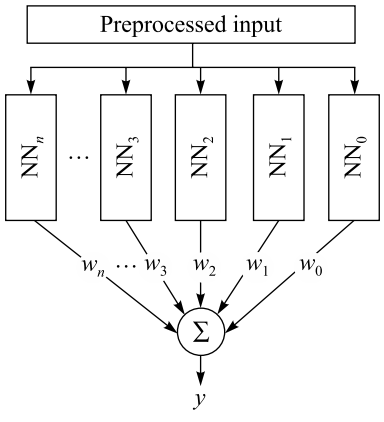

Fig. 1. Weighted ensemble

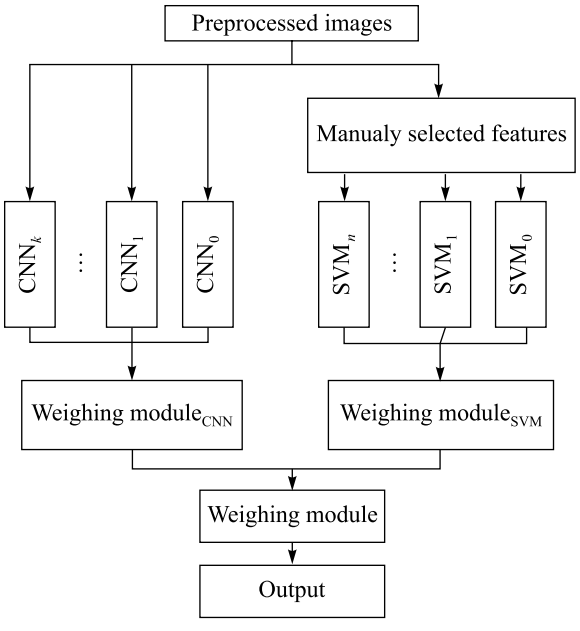

Fig. 2. Ensemble of CNN and SVM models

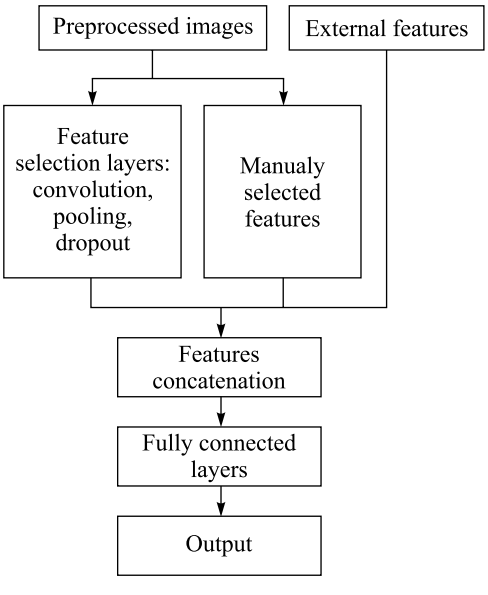

Fig. 3. Hybrid convolutional neural network

At the first stage, the first branch of the network is trained until sufficient accuracy is achieved, or before stopping by early stopping methods. At the second stage, the weights of the convolutional layers of the network are fixed and the training is carried out only for fully connected layers, where the features from convolutional layers, the manual set of features and the external features come together.

The proposed technique based on an ensemble of models for identifying objects of remote sensing of the Earth consists of the following steps.

Step 1: description of source data, the objects for identification and model quality measures (problem statement).

Step 2: formation of a training set: data collection, preprocessing, marking of the output set.

Step 3: searching additional features to solve the problem.

Step 4: expansion of the training set with additional features.

Step 5: splitting the training set into training and test sets.

Step 6: determination of the model's architecture based on the source data.

Step 7: determination of the hyperparameters range of the selected architecture.

Step 8: determination of the optimal model hyperparameters by grid or random search [14] using the *k*-fold cross-validation on the training set.

Step 9: forming an ensemble based on cross-validated models.

Step 10: if the model satisfies the quality measure on the test set, then the problem is solved, otherwise, it is necessary to expand the data set and go to step 4.

Experiments

The experimental data are images obtained using synthetic aperture radar (SAR), which allows taking radar images of the Earth's surface and objects on it, regardless of meteorological conditions and the level of the natural light of the area under observation. They include [15]:

• the images in two polarisation modes: horizontal – horizontal (HH), horizontal – vertical (HV); each image contains one object: a ship or an iceberg (fig. 4);

• incidence angle;

• data set: 1604 images with the size 75 × 75.

Data set was divided into 80 % training part and 20 % test part, so we use 1283 samples for training ENN.

Additionally, experiments were carried out on an extended data set. A simple augmentation technique was used for this: horizontal flip, vertical flip, 90-degree clockwise rotation, horizontal flip and vertical flip for the rotated image. In this case, we have 7698 samples for training.

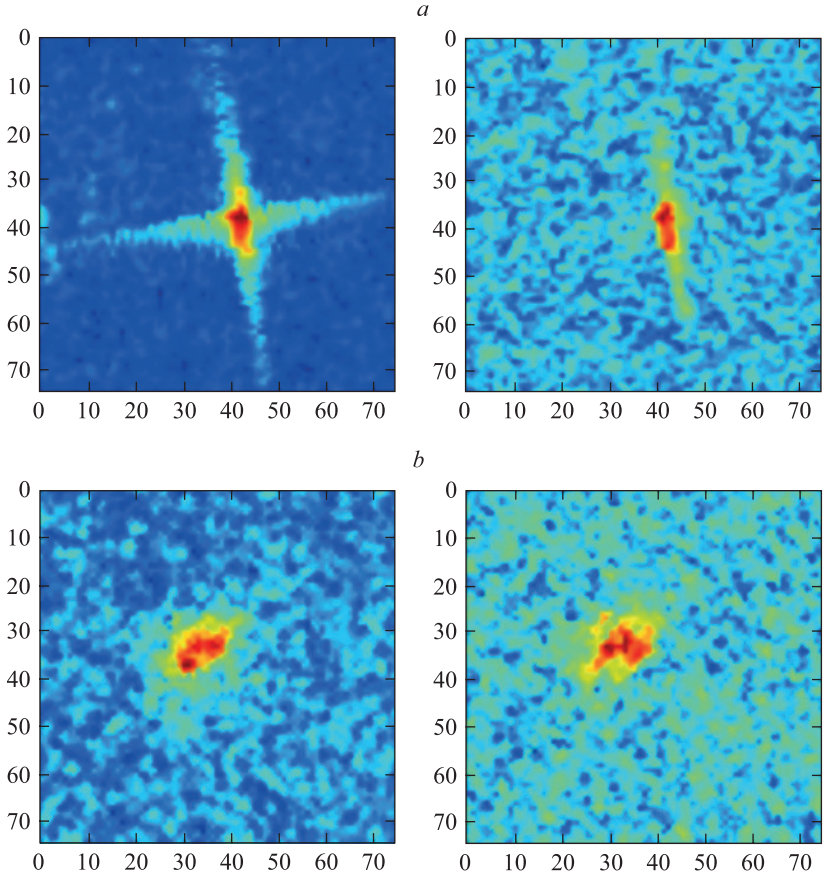

Fig. 4. Sample images: *a* – ship; *b* – iceberg

Training part was used for cross validated grid search, and test part was used for evaluation.

The task is to identify objects of two classes: an iceberg or a ship, which is essentially a binary classification task.

Efficiency in the classification problem can be assessed using accuracy – this is a basic measure that shows the proportion of correct model responses. For the binary classification problem, when the model derives the class probabilities, the logarithmic loss (logloss) function is used:

$$L = -\frac{1}{l}\sum_{i=1}^{l}\left(y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\right),$$

where $\hat{y}_i$ is a model response on the i object; y is a true class label on the i sample; l is the number of samples.

The minimisation of L can be represented as the task of maximising accuracy by a penalty for incorrect predictions. However, it should be noted that L is heavily penalised for the classifier's confidence in the wrong answer. Therefore, an error on one object can give a significant increase in the total error. Such samples are often outliers, which must be filtered or treated separately.

Based on the analysis of the initial data, the following baseline CNN is proposed (network parameters were chosen empirically):

- size of the input layer: 75 × 75 × 3;
- 2D convolutional layer 1 consists of 64 kernels with size 3 × 3, the activation function is rectified linear unit (ReLU);
- pooling layer with 2D max pooling and pool size of 2 × 2;
- dropout with probability 0.3;
- 2D convolutional layer 2: 128 kernels with size 3 × 3, RELU;
- 2D max pooling 2 × 2;
- dropout with probability 0.3;
- 2D convolutional layer 3: 128 kernels with size 3 × 3, RELU;
- 2D max pooling 2 × 2;
- dropout with probability 0.3;
- 2D convolutional layer 4: 64 kernels with size 3 × 3, RELU;
- 2D max pooling 2 × 2;
- dropout with probability 0.3;
- fully connected layer of 1024 neurons, RELU;
- dropout with probability 0.5;
- fully connected layer of 512 neurons, RELU;
- dropout with probability 0.5;
- the output fully connected layer of two outputs with a softmax activation function.

The CNN accepts a pseudo image in which the first image channel is represented by the HH channel of the original data, the second channel is represented by the HV image channel, and the third image channel is represented by their composition in form of a normalised sum.

The CNN training is performed using the Adam stochastic gradient algorithm [16]. Cross-validation is performed for $k = 5$. Each model was trained for 60 epochs. The batch size is 32. Early stopping procedure [23] was used with patience equal to 10. Starting learning rate parameter was 1e-4. It was reduced by factor equal to 0.5 based on «reduce LR on Plateau» [24] algorithm with the patience equal to 7 (the patience means the number of epochs with no improvement after which the learning rate begins to reduce). Finally, the ensemble of five CNN models was formed.

In addition to the proposed ENN, single models and ensembles based on them were considered using the following widely used architectures: VGG16 [17], ResNet50 [19], EfficientNet-B0 [20], Xception [18], MobileNet-v2 [22], DenseNet-121 [21]. To build the ensemble base model convolutional layers were taken from each model and three fully connected layers were added.

During the experiments, architectures with a small number of weights were selected, since the input data has a low dimension and a low detail. The ensembles of the best models from all trained models were also analysed.

For feature extraction, all input images were binarised using manually selected threshold. After this operation there was a mask for target objects on each image that presented any pseudo image. Also 61 features extracted from the pseudo images:

- 10 moments of the 1st and 2nd order calculated for both HH and HV images that describe an object shape;
- global statistics (mean, maximum, minimum, variance for both HH and HV images);
- differences in global statistics (3 features for both HH and HV images);
- global statistics in the masked area (mean, maximum, minimum, variance, a sum for HH and HV images);
- local statistics (maximum local standard deviation, 6 maximum values differences, variance for both HH and HV images);
- incidence angle.

Using this set of features, the ensemble of SVM models with a non-linear Gaussian kernel is trained, for which the optimal parameters C (SVM hyperparameter) and γ (Gaussian kernel parameter) of the model were determined by random search. Also, the weighted ensemble of CNN and SVM models is formed. Additionally, a hybrid CNN model is formed (see fig. 3) that combines the features extracted by convolutional layers and the set of 61 manual features.

The result of the evaluation of these models is presented in table 1. The result of the evaluation of model ensembles on augmented data is presented in table 2.

Table 1

Model evaluation on Statoil/C-CORE Iceberg Classifier without augmentation

Model	Logloss	Accuracy	Parameters
Five-fold cross validated			
EfficientNet-B0	0.482	82.118	5 886 621
ResNet50	0.392	86.044	26 211 201
Xception	0.471	86.667	23 484 969
VGG16	0.599	61.246	15 765 313
MobileNet-v2	0.695	50.156	4 095 041
Baseline CNN	0.321	86.791	888 897
DenseNet-121	0.336	87.601	8 612 417
Five-fold ensemble			
EfficientNet-B0	0.306	87.539	29 433 105
ResNet50	0.267	87.850	131 056 005
Xception	0.286	90.031	117 424 845
VGG16	0.567	79.439	78 826 565
MobileNet-v2	0.695	50.156	20 475 205
Baseline CNN	0.279	86.916	4 444 485
DenseNet-121	0.240	89.408	43 062 085
Top-5 model ensemble	0.227	91.277	75 533 421
Top-10 model ensemble	0.226	90.966	160 489 110
Weighted ensemble by loss			
EfficientNet-B0	0.293	87.227	29 433 105
ResNet50	0.272	87.227	131 056 005
Xception	0.279	90.343	117 424 845
VGG16	0.454	81.308	78 826 565
MobileNet-v2	0.695	50.156	20 475 205
Baseline CNN	0.279	87.227	4 444 485
DenseNet-121	0.252	89.408	43 062 085
Top-5 model ensemble	0.228	91.900	75 533 421
Top-10 model ensemble	0.221	92.211	160 489 110
SVM ensemble	0.303	86.292	7241
Top-5 CNN + SVM ensemble	0.227	92.523	75 540 662
Hybrid CNN	0.248	90.654	905 025

Table 2

Model evaluation on Statoil/C-CORE Iceberg Classifier on augmented data

Model	Logloss	Accuracy	Parameters
Five-fold cross validated			
EfficientNet-B0	0.434	85.234	5 886 621
ResNet50	0.332	87.664	26 211 201

Ending table 2

Model	Logloss	Accuracy	Parameters
Xception	0.381	87.850	23 484 969
VGG16	0.672	54.330	15 765 313
MobileNet-v2	0.542	75.514	4 095 041
Baseline CNN	0.302	86.854	888 897
DenseNet-121	0.326	87.539	8 612 417
Five-fold ensemble			
EfficientNet-B0	0.279	89.408	29 433 105
ResNet50	0.249	89.097	131 056 005
Xception	0.256	90.966	117 424 845
VGG16	0.661	50.156	78 826 565
MobileNet-v2	0.366	87.227	20 475 205
Baseline CNN	0.277	88.162	4 444 485
DenseNet-121	0.243	90.654	43 062 085
Top-5 model ensemble	0.243	90.966	90 405 973
Top-10 model ensemble	0.241	90.031	155 490 078
Weighted ensemble by loss			
EfficientNet-B0	0.277	89.408	29 433 105
ResNet50	0.261	89.097	131 056 005
Xception	0.255	90.654	117 424 845
VGG16	0.596	70.405	78 826 565
MobileNet-v2	0.313	87.539	20 475 205
Baseline CNN	0.269	87.539	4 444 485
DenseNet-121	0.244	90.343	43 062 085
Top-5 model ensemble	0.252	89.719	90 405 973
Top-10 model ensemble	0.245	90.343	155 490 078

First of all, it should be noted that with the selected training parameters in the base data set, some models could not find a solution (see VGG16 and MobileNet-v2 in table 1). For MobileNet-v2, the situation is improved with a data set augmentation. The data set augmentation technique used for this task has increased the accuracy of single models and an ensemble of models of the same architecture.

As can be seen, the complication of the model architecture gives a slight increase in accuracy. At the same time, the number of weighting parameters of the model greatly is increased when using a more complex architecture.

The weighted ensemble makes the possibility to improve the accuracy of some models on the base data set. At the same time, on the extended data set, the accuracy of the ensemble either remained unchanged or slightly is decreased. It can be said that weighting improves the overall accuracy of the ensemble in the case when it can contain both too weak and strong models. Otherwise, when all models are about the same level, weighting does not improve the classification.

The SVM ensemble, as a classical approach, has shown low accuracy for this task in comparison with the ENNs. However, the ensemble of the CNN and the SVM models shows the highest accuracy on the test data set.

So, the combination of models of different architectures and training methods can significantly increase the efficiency of classification. At the same time, the amount of consumed resources also increases accordingly. For an ensemble of five models, the memory consumption is increased fivefold. And since the models learn independently, there is no way to use shared weights. Also, when each model is applied sequentially to the input data, the inference time is increased fivefold. It makes sense to use the models with low memory consumption in this case. Also, independent model training gives a possibility to produce parallel inference on the models without increasing time.

Conclusion

The technique based on an ensemble of models for identifying objects of Earth remote sensing images was proposed. It includes the following steps: preparing data, object feature extraction, creating base identification models, optimising the model's hyperparameters, construction of the ensemble, processing the data by the ensemble.

The technique was applied to binary clastering of images obtained by synthetic aperture radar. Evaluation of the proposed models on experimental data has showed that one of the effective ways to increase accuracy in machine learning tasks is to form an ensemble of heterogeneous models trained on different sets of input features.

References

1. Kim M, Choi W, Jeon Y, Liu L. A hybrid neural network model for power demand forecasting. *Energies*. 2019;12(5):931. DOI: 10.3390/en12050931.

2. Frankel A, Tachida K, Jones R. Prediction of the evolution of the stress field of polycrystals undergoing elastic-plastic deformation with a hybrid neural network model. *Machine Learning: Science and Technology*. 2020;1(3):035005. DOI: 10.1088/2632-2153/ab9299.

3. Liu H, Yang R, Wang T, Zhang L. A hybrid neural network model for short-term wind speed forecasting based on decomposition, multi-learner ensemble, and adaptive multiple error corrections. *Renewable Energy.* 2021;165:573–594. DOI: 10.1016/j.renene.2020.11.002.

4. Ma C, Du X, Cao L. Analysis of multi-types of flow features based on hybrid neural network for improving network anomaly detection. *IEEE Access.* 2019;7:148363–148380. DOI: 10.1109/ACCESS.2019.2946708.

5. Berkhahn S, Fuchs L, Neuweiler I. An ensemble neural network model for real-time prediction of urban floods. *Journal of hydrology*. 2019;575:743–754. DOI: 10.1016/j.jhydrol.2019.05.066.

6. Cheng B, Wu W, Tao D, Mei S, Mao T, Cheng J. Random cropping ensemble neural network for image classification in a robotic arm grasping system. *IEEE Transactions on Instrumentation and Measurement*. 2020;69(9):6795–6806. DOI: 10.1109/TIM.2020.2976420.

7. Large scale visual recognition challenge [Internet; cited 29.01.2021]. Available from: http://image-net.org/challenges/LSVRC/2016/results.

8. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. *Neural computation*. 1989;1(4):541–551.

9. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge: MIT Press; 2016. 781 p.

10. Parikh D, Polikar R. An ensemble-based incremental learning approach to data fusion. *IEEE Transactions on Systems, Man and Cybernetics*. *Part B: Cybernetics*. 2007;37(2):437450. DOI: 10.1109/TSMCB.2006.883873.

11. Marushko EE, Doudkin AA. Ensembles of neural networks for forecasting of time series of spacecraft telemetry. *Optical Memory and Neural Networks*. 2017;26(1):47–54. DOI: 10.3103/S1060992X17010064.

12. Kourentzes N, Barrow D, Crone S. Neural network ensemble operators for time series forecasting. *Expert Systems with Applications.* 2014;41(9):4235–4244. DOI: 10.1016/j.eswa.2013.12.011.

13. Vapnik V. *The nature of statistical learning theory.* 2nd edition. New York: Springer; 1999. 314 p.

14. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *Machine Learning Research*. 2012;13:281305.

15. Statoil/C-CORE iceberg classifier challenge. Data [Internet; cited 29.01.2021]. Available from: https://www.kaggle.com/c/statoil-iceberg-classifier-challenge/data.

16. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv:1412.6980. 2017 [cited 29.01.2021]: [15 p.]. Available from: https://arxiv.org/abs/1412.6980.

17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [Preprint]. 2015 [cited 29.01.2021]: [14 p.]. Available from: https://arxiv.org/abs/1409.1556.

18. Chollet F. Xception: deep learning with depthwise separable convolutions. In: IEEE Computer Society. *2017 IEEE Conference on computer vision and pattern recognition (CVPR); 2017 July 21–26; Honolulu, USA.* Los Alamitos: IEEE; 2017. p. 1251–1258. DOI: 10.1109/CVPR.2017.195.

19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Computer Society. *Proceedings of the IEEE conference on computer vision and pattern recognition; 2016 June 27–30; Las Vegas, Nevada*. Los Alamitos: IEEE; 2016. p. 770–778. DOI: 10.1109/CVPR.2016.90.

20. Tan M, Le QV. Efficient net: rethinking model scaling for convolutional neural networks. arXiv:1905.11946 [Preprint]. 2020 [cited 29.01.2021]: [11 p.]. Available from: https://arxiv.org/abs/1905.11946.

21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: IEEE Computer Society. *2017 IEEE Conference on computer vision and pattern recognition (CVPR); 2017 July 21–26; Honolulu, USA.* Los Alamitos: IEEE; 2017. p. 2261–2269. DOI: 10.1109/CVPR.2017.243.

22. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobilenetV2: inverted residuals and linear bottlenecks. In: IEEE Computer Society. *2018 IEEE/CVF Conference on computer vision and pattern recognition; 2018 June 18–23; Salt Lake City, USA.* Los Alamitos: IEEE; 2018. p. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

23. Prechelt L. Early stopping – but when? In: Orr GB, Müller K-R, editors. *Neural Networks: tricks of the trade*. Berlin: Springer; 1998. p. 55–69.

24. Goyal P, Dollar P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, et al. Accurate, large minibatch SGD: training imagenet in 1 hour. arXiv:1706.02677 [Preprint]. 2018 [cited 29.01.2021]: [12 p.]. Available from: https://arxiv.org/abs/1706.02677.

Received 03.02.2021 / revised 29.06.2021 / accepted 29.06.2021.

АННОТАЦИИ ДЕПОНИРОВАННЫХ В БГУ РАБОТ
INDICATIVE ABSTRACTS OF THE PAPERS DEPOSITED IN BSU

УДК 517.958(075.8)
Козловская И. С. **Уравнения математической физики** [Электронный ресурс] : электрон. учеб.-метод. комплекс для спец.: 1-31 03 04 «Информатика»; 1-98 01 01 «Компьютерная безопасность (по направлениям)», направление спец. 1-98 01 01-01 «Компьютерная безопасность (математические методы и программные системы)» / И. С. Козловская ; БГУ. Электрон. текстовые дан. Минск : БГУ, 2021. 149 с. : ил. Библиогр.: с. 148–149. Режим доступа: https://elib.bsu.by/handle/123456789/257012. Загл. с экрана. Деп. в БГУ 15.03.2021, № 002915032021.

Электронный учебно-методический комплекс (ЭУМК) по учебной дисциплине «Уравнения математической физики» разработан в соответствии с образовательным стандартом первой ступени высшего образования для специальностей 1-31 03 04 «Информатика», 1-98 01 01 «Компьютерная безопасность (по направлениям)», направление специальности 1-98 01 01-01 «Компьютерная безопасность (математические методы и программные системы)», и предназначен для информационно-методического обеспечения преподавания дисциплины «Уравнения математической физики» для студентов данных специальностей. В ЭУМК содержится конспект лекций, перечень лабораторных занятий с материалами для работы, задания по управляемой самостоятельной работе.

УДК 517(075.8)
Мазаник С. А. **Математический анализ** [Электронный ресурс] : электрон. учеб.-метод. комплекс для спец. 1-31 03 04 «Информатика» : в 3 ч. Ч. 3 / С. А. Мазаник, О. А. Кастрица ; БГУ. Электрон. текстовые дан. Минск : БГУ, 2021. 105 с. : ил. Библиогр.: с. 94–97. Режим доступа: https://elib.bsu.by/handle/123456789/257817. Загл. с экрана. Деп. в БГУ 05.04.2021, № 003405042021.

Электронный учебно-методический комплекс (ЭУМК) по учебной дисциплине «Математический анализ» (часть 3) предназначен для студентов специальности 1-31 03 04 «Информатика». В ЭУМК содержится материал, изучаемый студентами второго курса в осеннем семестре. Содержание учебного материала включает разделы «Ряды», «Несобственные интегралы», «Интегралы, зависящие от параметра», «Ряды и интегралы Фурье», «Функции комплексной переменной». Изложение соответствует программе учебной дисциплины.

УДК 004.42:004.738.5(06)
Веб-программирование и интернет-технологии WebConf2021 [Электронный ресурс] : материалы 5-й Междунар. науч.-практ. конф. (Минск, 18–21 мая 2021 г.) / БГУ ; [редкол.: И. М. Галкин (отв. ред.) и др.]. Электрон. текстовые дан. Минск : БГУ, 2021. 400 с. : ил., табл. Библиогр. в тексте. Режим доступа: https://elib.bsu.by/handle/123456789/259432. Загл. с экрана. Деп. в БГУ 07.05.2021, № 005207052021.

Представлены тезисы и материалы докладов 5-й Международной научно-практической конференции «Веб-программирование и интернет-технологии WebConf2021», проводимой кафедрой веб-технологий и компьютерного моделирования механико-математического факультета Белорусского государственного университета. Тексты приведены в авторской редакции.

Адресовано преподавателям, студентам, аспирантам, разработчикам, занимающимся созданием и использованием веб-приложений и интернет-технологий.

УДК 517(075.8)+514.7(075.8)+512.623(075.8)
Математический анализ. Элементы дифференциальной геометрии. Теория поля. Теория функций комплексной переменной [Электронный ресурс] : электрон. учеб.-метод. комплекс для спец.: 1-31 04 02 «Радиофизика»; 1-31 04 03 «Физическая электроника»; 1-31 04 04 «Аэрокосмические радиоэлектронные и информационные системы и технологии»; 1-31 03 07 «Прикладная информатика (по направлениям)», направление спец. 1-31 03 07-02 «Прикладная информатика (информационные технологии телекоммуникационных систем)»; 1-98 01 01 «Компьютерная безопасность (по направлениям)», направление спец. 1-98 01 01-02 «Компьютерная безопасность (радиофизические методы и программно-технические средства)» / А. А. Егоров [и др.] ; БГУ. Электрон. текстовые дан. Минск : БГУ, 2021. 175 с. Библиогр.: с. 173–174. Режим доступа: https://elib.bsu.by/handle/123456789/261138. Загл. с экрана. Деп. в БГУ 08.06.2021, № 006408062021.

В электронном учебно-методическом комплексе (ЭУМК) по учебной дисциплине «Математический анализ. Элементы дифференциальной геометрии. Теория поля. Теория функций комплексной переменной» приводятся базовые понятия и краткие теоретические выкладки, дается решение большого числа типовых примеров и предлагается значительное количество задач с ответами к ним для самостоятельного решения. ЭУМК предназначен для студентов и преподавателей высших учебных заведений.

СОДЕРЖАНИЕ

ВЕЩЕСТВЕННЫЙ, КОМПЛЕКСНЫЙ И ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ

ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ И ОПТИМАЛЬНОЕ УПРАВЛЕНИЕ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ДИСКРЕТНАЯ МАТЕМАТИКА И МАТЕМАТИЧЕСКАЯ КИБЕРНЕТИКА

ВЫЧИСЛИТЕЛЬНАЯ МАТЕМАТИКА

ТЕОРЕТИЧЕСКАЯ И ПРИКЛАДНАЯ МЕХАНИКА

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

CONTENTS

Журнал включен Высшей аттестационной комиссией Республики Беларусь в Перечень научных изданий для опубликования результатов диссертационных исследований по физико-математическим наукам (в области математики и информатики), техническим наукам (в области информатики).

Журнал включен в наукометрические базы данных Scopus, Mathematical Reviews, Ulrichsweb, Google Scholar, zbMath, РИНЦ.

Журнал Белорусского государственного университета. Математика. Информатика. № 2. 2021

Учредитель:
Белорусский государственный университет

Юридический адрес: пр. Независимости, 4, 220030, г. Минск.
Почтовый адрес: пр. Независимости, 4, 220030, г. Минск.
Тел. (017) 259-70-74, (017) 259-70-75.
E-mail: jmathinf@bsu.by
URL: https://journals.bsu.by/index.php/mathematics

«Журнал Белорусского государственного университета. Математика. Информатика» издается с января 1969 г.
До 2017 г. выходил под названием «Вестник БГУ. Серия 1, Физика. Математика. Информатика» (ISSN 1561-834X).

Редакторы *О. А. Семенец, М. А. Подголина*
Технический редактор *В. В. Пишкова*
Корректор *Л. А. Меркуль*

Подписано в печать 30.07.2021.
Тираж 100 экз. Заказ 282.

Республиканское унитарное предприятие «Информационно-вычислительный центр Министерства финансов Республики Беларусь».
ЛП № 02330/89 от 03.03.2014.
Ул. Кальварийская, 17, 220004, г. Минск.

Journal of the Belarusian State University. Mathematics and Informatics. No. 2. 2021

Founder:
Belarusian State University

Registered address: 4 Niezaliežnasci Ave., Minsk 220030.
Correspondence address: 4 Niezaliežnasci Ave., Minsk 220030.
Tel. (017) 259-70-74, (017) 259-70-75.
E-mail: jmathinf@bsu.by
URL: https://journals.bsu.by/index.php/mathematics

«Journal of the Belarusian State University. Mathematics and Informatics» published since January, 1969.
Until 2017 named «Vestnik BGU. Seriya 1, Fizika. Matematika. Informatika» (ISSN 1561-834X).

Editors *O. A. Semenets, M. A. Podgolina*
Technical editor *V. V. Pishkova*
Proofreader *L. A. Merkul'*

Signed print 30.07.2021.
Edition 100 copies. Order number 282.

Republican Unitary Enterprise «Informatsionno-vychislitel'nyi tsentr Ministerstva finansov Respubliki Belarus'».
License for publishing No. 02330/89, 3 March 2014.
17 Kal'varyjskaja Str., Minsk 220004.