ISSN 1997-1397 (Print)
ISSN 2313-6022 (Online)

Журнал Сибирского федерального университета

Математика и физика

Journal of Siberian Federal University

Mathematics & Physics

2020 13 (2)

ISSN 1997-1997-1397
(Print)

ISSN 2313-6022
(Online)

2020 13 (2)

ЖУРНАЛ СИБИРСКОГО ФЕДЕРАЛЬНОГО УНИВЕРСИТЕТА
Математика и Физика

JOURNAL OF SIBERIAN FEDERAL UNIVERSITY
Mathematics & Physics

Издание индексируется Scopus (Elsevier), Emerging Sources Citation Index (WoS, Clarivate Analytics), Российским индексом научного цитирования (НЭБ), представлено в международных и российских информационных базах: Ulrich's periodicals directiory, ProQuest, EBSCO (США), Google Scholar, MathNet.ru, КиберЛенинке.

Включено в список Высшей аттестационной комиссии «Рецензируемые научные издания, входящие в международные реферативные базы данных и системы цитирования».

Все статьи представлены в открытом доступе http://journal.sfu-kras.ru/en/series/mathematics_physics.

Журнал Сибирского федерального университета.
Математика и физика.
Journal of Siberian Federal University. Mathematics & Physics.

Учредитель: Федеральное государственное автономное образовательное учреждение высшего образования "Сибирский федеральный университет"(СФУ)
Главный редактор: А.М. Кытманов. Редакторы: В.Е. Зализняк, А.В. Щуплев.
Компьютерная верстка: Г.В. Хрусталева
№ 2. 26.04.2020. Индекс: 42327. Тираж: 1000 экз. Свободная цена
Адрес редакции и издательства: 660041 г. Красноярск, пр. Свободный, 79, оф. 32-03.
Отпечатано в типографии Издательства БИК СФУ
660041 г. Красноярск, пр. Свободный, 82а.
Свидетельство о регистрации СМИ ПИ № ФС 77-28724 от 27.06.2007 г.,
выданное Федеральной службой по надзору в сфере массовых коммуникаций, связи и охраны культурного наследия
http://journal.sfu-kras.ru
Подписано в печать 15.04.20. Формат 84×108/16. Усл.печ. л. 11,9.
Уч.-изд. л. 11,6. Бумага тип. Печать офсетная.
Тираж 1000 экз. Заказ 11246
Возрастная маркировка в соответствии с Федеральным законом № 436-ФЗ:16+

CONTENTS

СОДЕРЖАНИЕ

DOI: 10.17516/1997-1397-2020-13-2-135-140
УДК 517.5

On the Application of the Plan Formula to the Study of the Zeta-Function of Zeros of Entire Function

Vyacheslav I. Kuzovatov*
Alexander M. Kytmanov†
Siberian Federal University
Krasnoyarsk, Russian Federation
Azimbai Sadullaev‡
National university of Uzbekistan
Tashkent, Uzbekistan

Received 10.09.2019, received in revised form 16.11.2019, accepted 20.01.2020

Abstract. We consider an application of the Plan formula to the study of the properties of the zeta-function of zeros of entire function. Based on this formula, we obtained an explicit expression for the kernel of the integral representation of the zeta-function in this case.

Keywords: zeta-function of zeros, Plan formula, integral representation.

Citation: V.I.Kuzovatov, A.M.Kytmanov, A.Sadullaev, On the Application of the Plan Formula to the Study of the Zeta-Function of Zeros of Entire Function, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 135–140. DOI: 10.17516/1997-1397-2020-13-2-135-140.

Introduction

Recall (see, for example, [1, Chapter I, S. 1.9]) that the classical Plan formula has the form

$$\sum_{n=0}^{\infty} g(n) = \frac{1}{2} g(0) + \int_0^{\infty} g(\tau)\, d\tau + i \int_0^{\infty} \frac{g(it) - g(-it)}{e^{2\pi t} - 1}\, dt, \tag{1}$$

and is valid if

1. $g(\zeta)$ is regular for $\operatorname{Re} \zeta \geqslant 0$, $\zeta = \tau + it$,
2. $\lim\limits_{t \to \infty} e^{-2\pi |t|} g(\tau + it) = 0$ uniformly for $0 \leqslant \tau < \infty$,
3. $\lim\limits_{\tau \to \infty} \int\limits_{-\infty}^{\infty} e^{-2\pi |t|} \left| g(\tau + it) \right| dt = 0$.

The Plan formula has been known for quite some time in the theory of functions of a complex variable. It is used in investigating analytical properties of functions assigned in the form of progressions and for finding the sums of progressions in final form. Various generalizations of the Plan formula are obtained in works [2–4].

*kuzovatov@yandex.ru
†AKytmanov@sfu-kras.ru https://orcid.org/0000-0002-7394-1480
‡sadullaev@mail.ru

Among the physical applications of the classical Plan formula (some of its generalizations) one may note in the theory of quantum fields for renormalizing the tensor of the energy pulse of a scalar field in different Friedmann models of the Universe. It is also used for calculating a vacuum mean tensor of the energy-pulse of quantum fields in different complete and incomplete manifolds (the Casimir effect). A detailed presentation of these issues may be found in work [2].

In this article we consider an application of the Plan formula to the study of the properties of the zeta-function of zeros of entire function.

Regarding generalizations of the zeta-function, we note that in 1950s I. M. Gel'fand, B. M. Levitan, and L. A. Dikii (see, for example, [5–7]) studied the zeta-function associated to eigenvalues of the Sturm-Liouville operator. As it turned out, its value is connected with the trace of the operator. Their approach was further developed by V. B. Lidskii and V. A. Sadovnichii [8] who considered a class of entire functions of one variable, defined the zeta-function of their zeroes and investigated its domain of analytic continuation. S. A. Smagin and M. A. Shubin [9] constructed the zeta-functions for elliptic operators, as long as for operators of more general type, proved a possibility of meromorphic continuation of the zeta-function and gave some information on its poles.

Multidimensional results were obtained by A. M. Kytmanov and S. G. Myslivets [10]. They introduced the concept of the zeta-function associated with a system of meromorphic functions $f = (f_1, \ldots, f_n)$ in $\mathbb{C}^n$. Using the residue theory, these authors gave an integral representation for the zeta-function, but the system of functions $f_1, \ldots, f_n$ was subject to rigid constraints.

1. Auxiliary results

Let $f(z)$ be an entire function of order ρ in $\mathbb{C}$. Consider the equation

$$f(z) = 0. \tag{2}$$

Denote by $N_f = f^{-1}(0)$ the set of all solutions to (2) (we take every zero as many times as its multiplicity). The numbers of roots is at most countable.

The zeta-function $\zeta_f(s)$ of Eq. (2) is defined in the following way:

$$\zeta_f(s) = \sum_{z_n \in N_f} (-z_n)^{-s},$$

where $s \in \mathbb{C}$.

In [11], using the residue theory, V. I. Kuzovatov and A. A. Kytmanov obtained two integral representation for the zeta-function constructed by zeros of an entire function of finite order on the complex plane. With the help of these representations, they described a domain which the zeta-function can be extended to.

Theorem 1.1 ([11]). *Let $f(z)$ be an entire function of the zero order in $\mathbb{C}$ and satisfy the condition*

$$\frac{f'(z)}{f(z)} - \omega_0 = O\left(\frac{1}{|z|}\right).$$

Suppose that $0 < Re\ s < 1$. Then

$$\zeta_f(s) = \frac{\sin \pi s}{\pi} \int_0^\infty \left(\frac{f'(x)}{f(x)} - \omega_0\right) x^{-s}\, dx, \tag{3}$$

where ω_0 is the limit value of $\dfrac{f'(x)}{f(x)}$ at infinity.

The method of proof of Theorem 1.1 shows that the statement remains valid in the case when $f(z)$ is an entire function of order less than 1.

Now we will give an integral representation for the zeta-function $\zeta_f(s)$ of zeros z_n of f which are $z_n = -q_n + is_n$, $q_n > 0$. Let us denote

$$F(f,x) = \sum_{n=1}^{\infty} e^{z_n x}. \tag{4}$$

We will assume that Re $s = \sigma > 1$ and the following conditions hold:

$$\varliminf_{n\to\infty} \frac{q_n}{n} > 0, \tag{5}$$

$$\text{the series} \quad \sum_{n=1}^{\infty} \left(\frac{1}{q_n}\right)^{\sigma-1} \text{converges.} \tag{6}$$

For the convergence of the series (4), using condition (5), it is necessary and sufficient (for real x) that $x > 0$ [11].

Theorem 1.2 ([11])**.** *Suppose that the conditions* (5) *and* (6) *are satisfied and Re* $s > 1$*. Then*

$$\zeta_f(s) = \frac{1}{\Gamma(s)} \int_0^{\infty} x^{s-1} F(f,x)\, dx,$$

where $F(f,x)$ *is defined by formula* (4)*, and* $\Gamma(s)$ *is the Euler gamma-function.*

Our goal is to obtain an explicit expression for the kernel of the integral representation (3) in case $z_n = -\pi n^2$. This choice of zeros z_n is due to the fact that for series

$$F(f,x) = \sum_{n=1}^{\infty} e^{z_n x} = \sum_{n=1}^{\infty} e^{-\pi n^2 x} := \psi(x)$$

for $x > 0$ it is known (see, for example, [12, Chapter II, S. 6]) that

$$2\psi(x) + 1 = \frac{1}{\sqrt{x}} \left\{ 2\psi\left(\frac{1}{x}\right) + 1 \right\}.$$

2. The main result

Theorem 2.1. *Let* $f(z)$ *be an entire function of order* ρ *with zeros* $z_n = -\pi n^2$*. Then for real* $x \in (0; +\infty)$ *the following holds*

$$\frac{f'(x)}{f(x)} = \frac{\sqrt{\pi}}{2\sqrt{x}} - \frac{1}{2x}.$$

Proof. We consider entire functions $f(z)$ of order ρ, which have the form

$$f(z) = C \prod_{n=1}^{\infty} \left(1 - \frac{z}{z_n}\right). \tag{7}$$

The representation (7) is true, for example, for entire functions of order less than 1 or for entire functions of the first order with the additional condition (the series $\sum\limits_{n=1}^{\infty} \frac{1}{|z_n|}$ is convergent). In particular, the representation (7) is true for functions of the zero genus.

It is easy to show that in this case we obtain

$$\frac{f'(z)}{f(z)} = \sum_{n=1}^{\infty} \frac{1}{z - z_n} \tag{8}$$

if $z \neq z_n$.

Since the order of the canonical product (7) is equal to the index of convergence ρ_1 of its zeros and for given values of z_n

$$\rho_1 = \overline{\lim_{n\to\infty}} \frac{\ln n}{\ln |z_n|} = \frac{1}{2},$$

then representations (7) and (8) are true for considered function $f(z)$.

Now we make use of a summation formula due to Plan (1). Taking

$$g(\zeta) = \frac{1}{z + \pi\zeta^2}, \quad \text{Re } z > 0,$$

in (1) we find that

$$\sum_{n=1}^{\infty} \frac{1}{z + \pi n^2} = -\frac{1}{2z} + \int_0^{\infty} \frac{d\tau}{z + \pi\tau^2}.$$

Passing in the last equality from complex z to real $x \in (0; +\infty)$, we obtain

$$\sum_{n=1}^{\infty} \frac{1}{x + \pi n^2} = \frac{f'(x)}{f(x)} = -\frac{1}{2x} + \int_0^{\infty} \frac{d\tau}{x + \pi\tau^2} = -\frac{1}{2x} + \frac{1}{\pi}\int_0^{\infty} \frac{d\tau}{\tau^2 + x/\pi} =$$

$$= -\frac{1}{2x} + \frac{1}{\pi}\sqrt{\frac{\pi}{x}}\,\operatorname{arctg}\frac{\tau\sqrt{\pi}}{\sqrt{x}}\bigg|_0^{\infty} = -\frac{1}{2x} + \frac{1}{\sqrt{\pi x}}\cdot\frac{\pi}{2} = \frac{\sqrt{\pi}}{2\sqrt{x}} - \frac{1}{2x}. \qquad \square$$

Corollary 1. *Suppose that the conditions of Theorem 2.1 are satisfied. If ω_0 is the limit value of $\dfrac{f'(x)}{f(x)}$ at infinity, i.e.*

$$\omega_0 = \lim_{x\to+\infty} \frac{f'(x)}{f(x)},$$

then $\omega_0 = 0$.

Remark 1. If f is an arbitrary entire function of order $1 \leqslant \rho < \infty$, with zeros $z_n = -\pi n^2$, then the ratio can be represented as

$$\frac{f(z)}{\prod\limits_{n=1}^{\infty}\left(1 - \frac{z}{z_n}\right)} = e^{g(z)},$$

where $g(z)$ is an entire function. Since $1 \leqslant \rho < \infty$, $g(z)$ is a polynomial, $\deg g = \rho$, and $\rho \in \mathbb{N}$ [13]. Therefore,

$$f(z) = \Pi(z)e^{g(z)}, \quad \Pi(z) = \prod_{n=1}^{\infty}\left(1 - \frac{z}{z_n}\right),$$

and

$$\frac{f'(z)}{f(z)} = \frac{\Pi'(z)e^{g(z)} + \Pi(z)e^{g(z)}g'(z)}{\Pi(z)e^{g(z)}} = \frac{\Pi'(z)}{\Pi(z)} + g'(z).$$

Consequently in this case we take

$$\frac{f'(x)}{f(x)} = \frac{\sqrt{\pi}}{2\sqrt{x}} - \frac{1}{2x} + g'(x), \quad 1 \leqslant \rho < \infty.$$

The work was supported by RFBR, grant 18-51-41011 Uzb.t and grant 18-31-00019.

References

[1] H.Bateman, A.Erdelyi, Higher Transcendental Functions. V. 1, New York, MC Graw-Hill Book Company, 1953.

[2] A.A.Saaryan, Towards an Abel-Plana Summing Formula, *Soviet Journal of Contemporary Physics*, **21**(1986), no. 5, 32–36.

[3] V.I.Kuzovatov, A.M.Kytmanov, *Journal of Contemporary Mathematical Analysis*, **53**(2018), no. 3, 139–146. DOI: 10.3103/S1068362318030044

[4] V.I.Kuzovatov, Generalization of the Plana Formula, *Russian Mathematics*, **62**(2018), no. 5, 34–43.

[5] I.M.Gel'fand, B.M.Levitan, On a Simple Identity for Eigenvalues of a Second Order Differential Operator, *Sov. Phys. Dokl.*, **88**(1953), no. 4, 593–596 (in Russian).

[6] L.A.Dikii, The Zeta-Function of an Ordinary Differential Equation on a Finite Interval, *Izv. Akad. Nauk SSSR Ser. Mat.*, **19**(1955), no. 4, 187–200 (in Russian).

[7] L.A.Dikii, Trace Formulas for Sturm-Liouville Differential Operators, *Uspekhi Mat. Nauk*, **13**(1958), no. 3, 111–143 (in Russian).

[8] V.B.Lidskii, V.A.Sadovnichii, Regularized Sums of Zeros of a Class of Entire Functions, Functional Analysis and Its Applications, *Funct. Anal. Appl.*, **1**(1967), no. 2, 133–139.

[9] S.A.Smagin, M.A.Shubin, On the Zeta-Function of a Transversally Elliptic Operator, *Russian Mathematical Surveys*, **39**(1984), no. 2, 201–202.

[10] A.M.Kytmanov, S.G.Myslivets, *Siberian Math. J.*, **48**(2007), no. 5, 863–870. DOI: 10.1007/s11202-007-0088-z

[11] V.I.Kuzovatov, A.A.Kytmanov, On the Zeta-Function of Zeros of Some Class of Entire Functions, *J. Siberian Federal Univ. Math. Phys.*, **7**(2014), no. 4, 489–499.

[12] E.C.Titchmarsh, The Theory of the Riemann Zeta-Function, Oxford University Press, Oxford, 1951.

[13] A.Sadullaev, On the canonical decomposition of entire functions, *Theory of functions, functional analysis and their applications,* (1974), no. 21, 107–121.

О применении формулы Плана к исследованию дзета-функции нулей целой функции

Вячеслав И. Кузоватов
Александр М. Кытманов
Сибирский федеральный университет
Красноярск, Российская Федерация
Азимбай Садуллаев
Национальный университет Узбекистана
Ташкент, Узбекистан

Аннотация. В данной статье рассмотрено применение формулы Плана к исследованию свойств дзета-функции нулей целой функции. На основе данной формулы получено явное выражение для ядра интегрального представления дзета-функции в этом случае.

Ключевые слова: дзета-функция нулей, формула Плана, интегральное представление.

DOI: 10.17516/1997-1397-2020-13-2-141-150
УДК 517.9

Root Locus of Algebraic Equations

Valeriy I. Bykov*
Svetlana B. Tsybenova†
Emanuel Institute of Biochemical Physics RAS
Moscow, Russian Federation

Received 28.02.2017, received in revised form 18.10.2018, accepted 15.01.2020

Abstract. The locus of real and complex roots of algebraic equations are constructed in this paper. Calculations of specific equations show that the location of their roots depends on the type of equation.

Keywords: algebraic equation, real root, complex root, root locus, nonlinear equation with parameters.

Citation: V.I.Bykov, S.B.Tsybenova, Root Locus of Algebraic Equations, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 141–150. DOI: 10.17516/1997-1397-2020-13-2-141-150.

Introduction

In this paper, we consider some of specific algebraic equations. However, our experience shows that the reported examples allow us to understand the structure and features of the geometric location of the roots under changes of parameters in the general case.

Systems of nonlinear equations with parameters arise, for example, in the analysis of the stability of stationary equations of chemical kinetics or in mathematical models of chemical technologies [1–4]. From a formal point of view, parametric analysis of specific mathematical model leads to the study of the geometric location of roots of nonlinear (including algebraic) equations [5–11].

In this paper we consider specific, but very important, case of an algebraic equation of the form:

$$F(z,\, p) = 0, \tag{1}$$

where function F is a polynomial, z is unknown variable. p is a parameter which is varied through a wide rage: $-\infty < p < +\infty$. As parameter p changes $z(p)$ defined by equation (1) can continuously change (deformation) or change abruptly (reconstruction, bifurcation). Thom's classification theorem [12,13] gives enumeration of possible local structures at bifurcation points. However, it is often important to know the quantitative characteristics of possible deformations of the solutions z under changes of parameters p in a wide interval. For example, it is important to determine not only the stability of steady states but also the degree of their stability and the stability margin of phase [2,3].

*vibykov@mail.ru https://orcid.org/0000-0003-0775-385X
† https://orcid.org/0000-0001-5599-0580

1. General analysis

The main problem of the qualitative and numerical analysis of systems of nonlinear equations with parameter (1) is to construct the root locus $z(p)$. As a rule, the problem to construct the inverse relation

$$p = p(z). \tag{2}$$

is a more simple problem.

Construction of the root locus of algebraic equation

$$a_n z^n + a_{n-1} z^{n-1} + ... + a_1 z + a_0 = 0, \tag{3}$$

is considered, where a_i are real coefficients. Without loss of generality, instead of (3) the following algebraic equation

$$z^n + a_{n-1} z^{n-1} + ... + a_1 z + 1 = 0. \tag{4}$$

is considered. Taking one of the coefficients in (4) as parameter p, rewrite this equation in the following way

$$p = z^{n-k} + ... + \frac{1}{z^k}, \tag{5}$$

where $n < k < 1$ and $z \in (-\infty, +\infty)$. For each fixed k, parameter p is a function of real variable z. Relation (5) allows us to determine the inverse function

$$z = \varphi(p). \tag{6}$$

Relation (5) allows us to relatively easy obtain $\varphi(p)$ which is the locus of real roots of original algebraic equation (4). The locus of complex roots z of this equation under changes of parameter p can be constructed from the condition

$$\text{Im } p(z) = 0. \tag{7}$$

The simplest problem for equation (3) is to construct the locus of real and complex roots of algebraic equations with three monomials

$$z^n - p z^{n-k} + 1 = 0, \tag{8}$$

where $1 < k < n$; n and k are integers. Let us note that quadratic and cubic equations without loss of generality quite fit in case (8). From equation (8) we have

$$p = z^k + \frac{1}{z^{n-k}}. \tag{9}$$

Relation $p(z)$ defined in (9) specifies the locus of real roots $z(p)$ or rather its inverse function $p(z)$, where z is real variable.

Differentiating (9) with respect to z, we obtain conditions for critical values of the parameter:

$$\frac{dp}{dz} = k z^{k-1} - \frac{n-k}{z^{n-k+1}} = 0, \tag{10}$$

from here

$$z_* = \big(\frac{n-k}{k}\big)^1 / n, \tag{11}$$

$$p_* = (z_*)^k + \frac{1}{z_*^{n-k}}. \tag{12}$$

In accordance with the general scheme of studying of (5)–(7), the locus of complex roots of equation (8) on the complex plane is defined according to (9), taking into account the condition $\text{Im } p(z) = 0$.

2. Specific examples

Implementation of the general scheme of construction of the root locus of algebraic equation (8) is illustrated by some specific examples.

Let us start with the simplest quadratic equation

$$z^2 - pz + 1 = 0. \tag{13}$$

From (13) we have the locus for real roots

$$p = z + \frac{1}{z}. \tag{14}$$

Relation (14) is shown in Fig. 1. The critical value of parameter p_* is

$$p_* = 2, \quad z_* = 1. \tag{15}$$

From the condition Im $p = 0$ for (14) we have: $a^2 + b^2 = 1$ on the complex plane $(a,\ b)$, i.e., for various p complex roots lie on the unit circle (see Fig. 2). Let us consider a cubic equation

$$z^3 - pz + 1 = 0. \tag{16}$$

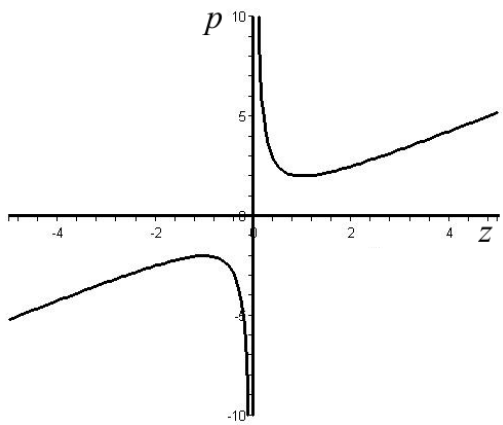

Fig. 1. Locus of real roots of equation (13)

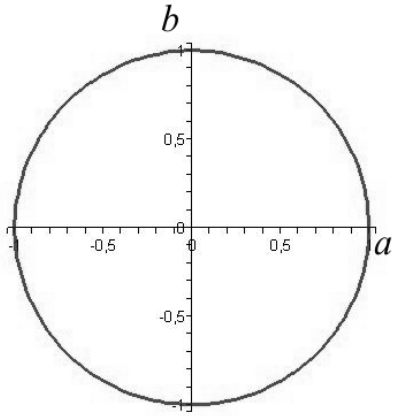

Fig. 2. Locus of complex roots $a^2 + b^2 = 1$

It is easy to show that any cubic equation

$$a_3x^3 + a_2x^2 + a_1x + a_0 = 0$$

is reduced to (16), using shift and stretch of argument x. Equation (16) contains a single parameter p. From (16) we have

$$p = z^2 + \frac{1}{z}. \tag{17}$$

Relation (17) defines the locus of real roots of cubic equation (16) (see Fig. 3). Critical values of the parameter and variable are

$$p_* = \sqrt[3]{2} + \frac{1}{\sqrt[3]{4}}, \quad z_* = \frac{1}{\sqrt[3]{2}}. \tag{18}$$

It follows from the condition Im $p = 0$ and (17) that locus of complex roots of equation (16) on the complex plane $(a,\ b)$ is

$$a^2 + b^2 = \frac{1}{2a}. \tag{19}$$

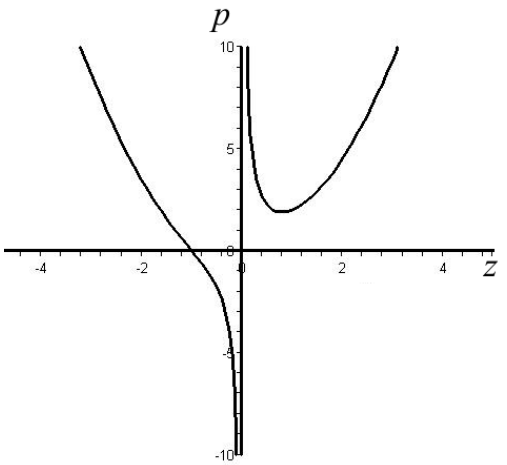

Fig. 3. Locus of real roots of equation (16)

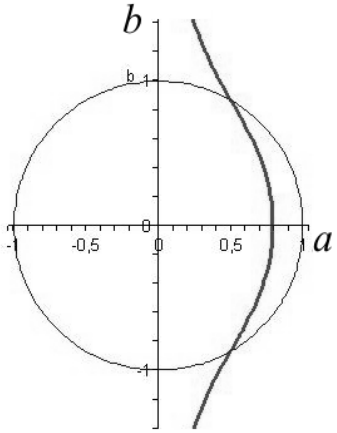

Fig. 4. Locus of complex roots $a^2+b^2=1/2a$

At $b=0$ we have $a=1/\sqrt[3]{2}$. Relation (19) is shown in Fig. 4.

For the cubic equation

$$z^3 - pz^2 + 1 = 0 \tag{20}$$

we have

$$p = z + \frac{1}{z^2}. \tag{21}$$

The loci of real and complex roots of (20) are presented in Fig. 5 and Fig. 6. On the complex plane the roots of equation (21) lie on the curve

$$(a^2+b^2)^2 = 2a. \tag{22}$$

The critical values p_* and z_* are defined in (18).

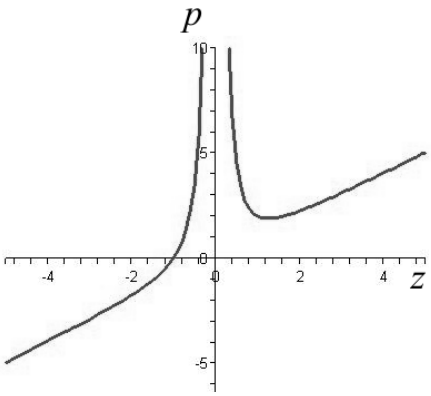

Fig. 5. Locus of real roots of equation (20)

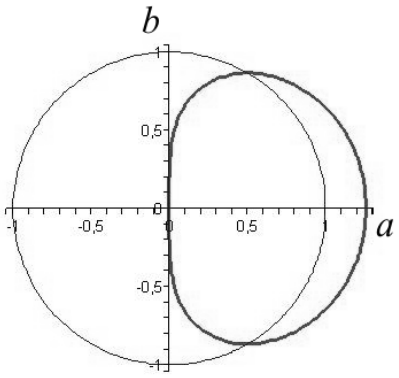

Fig. 6. Locus of complex roots (22)

Let us consider several variants of the 4th degree equation. For equation

$$z^4 - pz + 1 = 0 \tag{23}$$

real roots are specified by relation

$$p = z^3 + \frac{1}{z}. \tag{24}$$

Complex roots are defined as above, using condition $Im\ p = 0$:

$$(3a^2 - b^2)(a^2 + b^2) = 1. \tag{25}$$

Relations (24) and (25) are illustrated in Fig. 7 and Fig. 8, where

$$p_* = \pm\left(\sqrt[4]{3} + \frac{1}{\sqrt[3]{81}}\right), \quad z_* = \pm\frac{1}{\sqrt[4]{3}}. \tag{26}$$

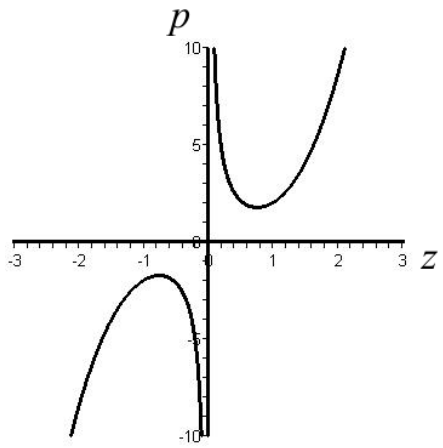

Fig. 7. Locus of real roots of equation (23)

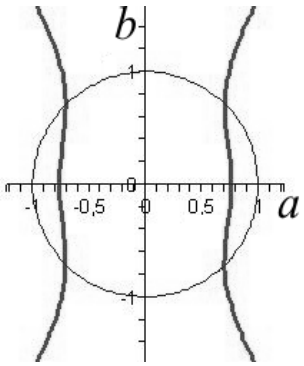

Fig. 8. Locus of complex roots (25)

For the 4th degree equation of the form

$$z^4 - pz^2 + 1 = 0 \tag{27}$$

we have

$$p = z^2 + \frac{1}{z^2}; \quad \text{Im } p = 0: \ (a^2 + b^2)^3 = 1.$$

The loci of real and complex roots are shown in Fig. 9 and Fig. 10. In this case

$$p_* = 2, \quad z_* = \pm 1.$$

For the 4th degree equation of the form

$$z^4 - pz^3 + 1 = 0 \tag{28}$$

we have

$$p = z + \frac{1}{z^3}; \tag{29}$$

$$\text{Im } p = 0: \quad (a^2 + b^2)^3 = 3a^2 - b^2. \tag{30}$$

The loci of real and complex roots for (28) are shown in Fig. 11 and Fig. 12. Here critical values z_* and p_* have the form

$$p_* = \pm\left(\sqrt[4]{3} + \frac{1}{\sqrt[4]{27}}\right), \quad z_* = \pm\sqrt[4]{3}.$$

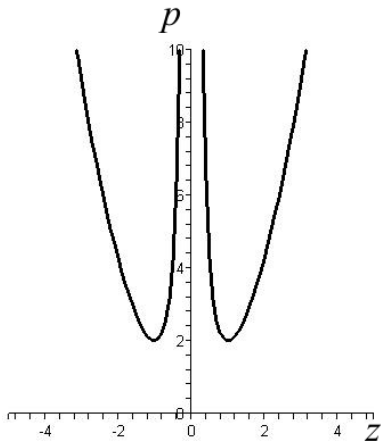

Fig. 9. Locus of real roots of equation (27)

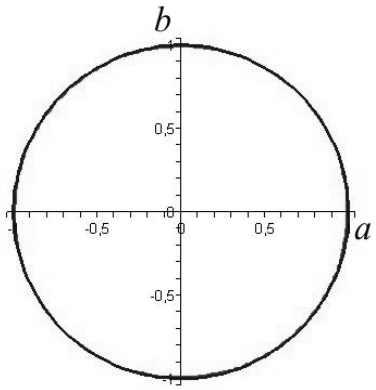

Fig. 10. Locus of complex roots $(a^2+b^2)^2 = 1$

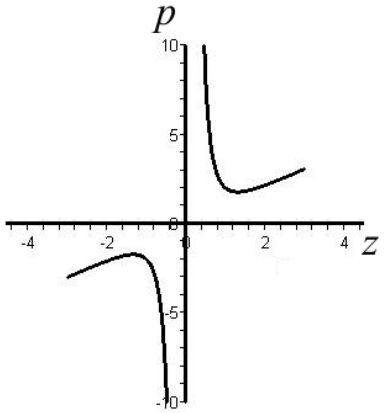

Fig. 11. Locus of real roots of equation (28)

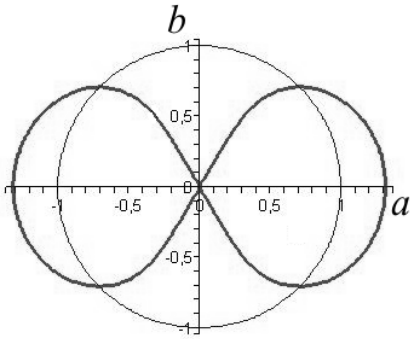

Fig. 12. Locus of complex roots (30)

Consider equations of higher degree. For the 5th degree equation

$$z^5 - pz + 1 = 0 \tag{31}$$

we have

$$p = z^4 + \frac{1}{z}; \tag{32}$$

$$\text{Im } p = 0: \quad 4a(a^4 - b^4) = 1. \tag{33}$$

The loci of real and complex roots of (32) and (33) are shown in Fig. 13 and Fig. 14. Critical values p_* and z_* are

$$p_* = \sqrt[5]{4} + \frac{1}{z_*^4}, \quad z_* = \sqrt[5]{4}.$$

The loci of real and complex roots of algebraic 5th degree equations

$$z^5 - pz^3 + 1 = 0 \tag{34}$$

and

$$z^5 - pz^4 + 1 = 0 \tag{35}$$

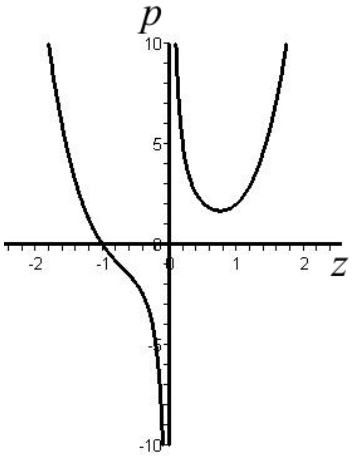

Fig. 13. Locus of real roots of equation (31)

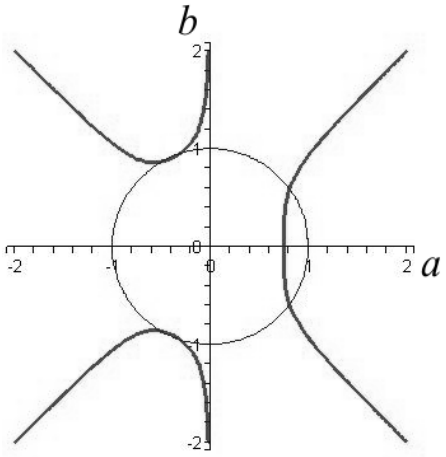

Fig. 14. Locus of complex roots (33)

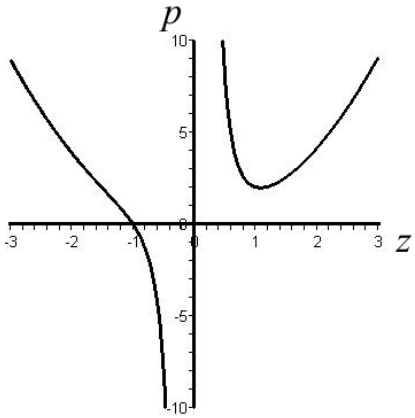

Fig. 15. Locus of real roots of equation (34)

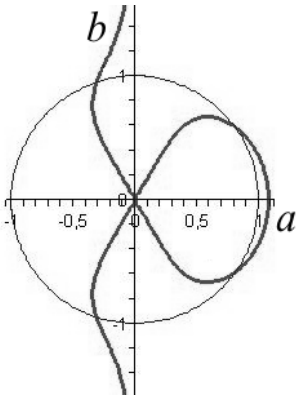

Fig. 16. Locus of complex roots of equation (34)

are presented in Figs. 15–18.

The loci of complex roots of equations of higher degree

$$z^8 - pz + 1 = 0 \tag{36}$$

and

$$z^8 - pz^7 + 1 = 0 \tag{37}$$

are shown in Figs. 19, 20.

Presented results allows us to estimate by analogy the qualitative nature of the corresponding parametric portraits of various types of algebraic equations of higher degree. For example, the loci of complex roots of equations

$$z^9 - pz^8 + 1 = 0 \tag{38}$$

and

$$z^{10} - pz^9 + 1 = 0 \tag{39}$$

are shown in Fig. 21 and 22.

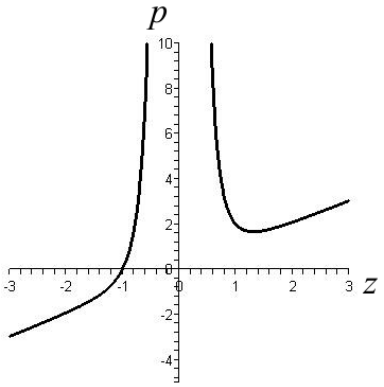

Fig. 17. Locus of real roots of equation (35)

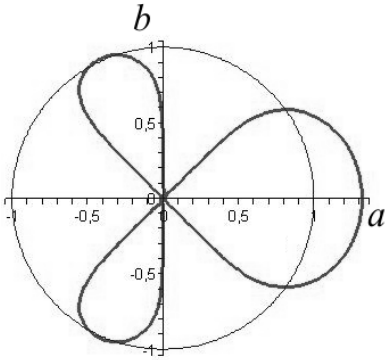

Fig. 18. Locus of complex roots of equation (35)

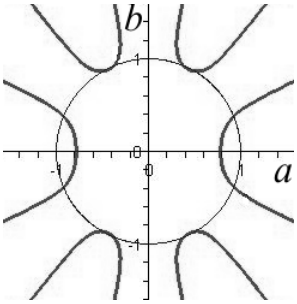

Fig. 19. Locus of real roots of equation (36)

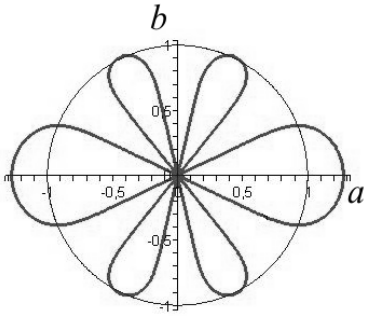

Fig. 20. Locus of complex roots of equation (37)

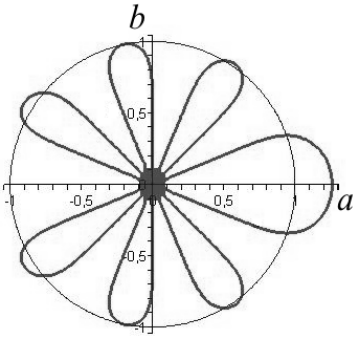

Fig. 21. Locus of complex roots of equation (38)

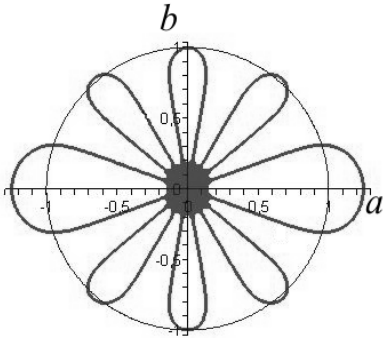

Fig. 22. Locus of complex roots of equation (39)

Conclusion

Comparative analysis of algebraic equations (38), (39) allows us to conclude that complex roots of equations

$$z^m - pz + 1 = 0 \tag{40}$$

at large m, rather densely fill the whole complex plane except the inside of unit circle with the centre at the origin. On the contrary, for equations

$$z^m - pz^{m-1} + 1 = 0 \tag{41}$$

complex roots are located inside of unit circle with the centre at the origin. For equations

$$z^m - pz^{m-k} + 1 = 0, \tag{42}$$

where k changes from 1 to $m-1$, the situation is intermediate between situations for (40) and (41). One part of the branches of the root locus on the complex plane goes to infinity, and other part is localized in a neighbourhood of the origin.

The performed analysis of several specific cases allows us to do some conclusions for algebraic equations of more general type

$$z^m + a_1 z^{m-1} + ... + a_{m-1} z + 1 = 0. \tag{43}$$

As real parameters a_i change, the loci of real roots of equation (43) can be easily determined by analogy with the performed analysis. To determine the locus of complex roots of equation (43) the results for some special cases presented in Figs. 9–22 will be undoubtedly useful.

Calculations show that a parametric portrait of an algebraic equation significantly depends on the type of an equation. Trinomial equations are simply analyzed. In typical case, the number of branches on which the real and complex roots are located is determined by the number of nonzero terms of this equation. The character of the localization of complex roots depends on the equation degree and degree of unknown variable at which the varying parameter is set.

Thus, the presented elementary numerical and qualitative analysis of specific algebraic equations allows us not only to get specific results of the analysis of the parameter dependence of equation roots but also to work out the understanding of qualitative features of curves on which roots of algebraic equations are located. It gives us the possibility to predict the qualitative behaviour of these roots as parameters change over a wide range.

References

[1] R.Aris, Introduction to the analysis of chemical reactors, Englewood Cliffs, NJ, USA, Prentice-Hall, 1965.

[2] V.I.Bykov, Modeling of critical phenomena in chemical kinetics, Moscow, URSS, 2014 (in Russian).

[3] V.I.Bykov, S.B.Tsybenova, Nonlinear models of chemical kinetics, Moscow, URSS, 2011 (in Russian).

[4] B.V.Volter, I.E.Sal'nikov, Stability of work regimes of the chemical reactors, Moscow, Khimiya, 1982 (in Russian).

[5] A.G. Kurosh, Course of higher algebra, Moscow, Nauka, 1965.

[6] V.I.Bykov, A.M.Kytmanov, M.Z.Lazman, Elimination methods in polynomial computer algebra, Novosibirsk, Nauka, 1991 (in Russian).

[7] V.I.Bykov, A.M.Kytmanov, M.Z.Lazman, M.Passare (ed.), Elimination methods in polynomial computer algebra, Netherlands, Springer Science and Business Media, 2012.

[8] B.P.Demidovich, Lectures on stability theory, Moscow, Nauka, 1967.

[9] F.Klein, Lectures on the icosahedron and the solution of equations of the fifth degree, NY, USA, Dover Publ., 1956.

[10] F.Klein, Elementary Mathematics from an Advanced Standpoint: Vol. I. Arithmetic, Algebra, Analysis, NY, USA, Dover Publ., 2004.

[11] F.R.Gantmakher, The theory of matrices, Moscow, Nauka, 1967.

[12] V.I.Arnold, Catastrophe theory, Moscow, Znanie, 1981.

[13] T.Poston, I.Stewart, Catastrophe theory and its applications, NY, USA, Dover Publ., 1996.

Геометрическое место корней алгебраических уравнений

Валерий И. Быков
Светлана Б. Цыбенова
Институт биохимической физики им. Н. М. Эмануэля РАН
Москва, Российская Федерация

Аннотация. В работе построено геометрическое место действительных и комплексных корней алгебраических уравнений. Расчеты конкретных уравнений показывают, что расположение этих корней существенно зависит от вида уравнения.

Ключевые слова: алгебраические уравнения, действительные и комплексные корни, геометрическое место корней, нелинейные уравнения с параметрами.

DOI: 10.17516/1997-1397-2020-13-2-151-159
УДК 517.9

Singular Quasilinear Elliptic Systems with (super-) Homogeneous Condition

Hana Didi*
Brahim Khodja†
Badji-Mokhtar Annaba University
Annaba, Algeria
Abdelkrim Moussaoui‡
A. Mira Bejaia University
Bejaia, Algeria

Received 02.10.2019, received in revised form 09.12.2019, accepted 16.01.2020

Abstract. In this paper we establish existence, nonexistence and regularity of positive solutions for a class of singular quasilinear elliptic systems subject to (super-) homogeneous condition. The approach is based on sub-supersolution methods for systems of quasilinear singular equations combined with perturbation arguments involving singular terms.

Keywords: singular system, p-Laplacian, sub-supersolution, regularity.

Citation: H.Didi, B.Khodja, A.Moussaoui, Singular Quasilinear Elliptic Systems with (super-) Homogeneous Condition, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 151–159.
DOI: 10.17516/1997-1397-2020-13-2-151-159.

Introduction

We consider the following system of quasilinear and singular elliptic equations:

$$(\mathcal{P}) \qquad \begin{cases} -\Delta_{p_1} u_1 = \lambda u_1^{\alpha_1} u_2^{\beta_1} + \delta h_1(x) \text{ in } \Omega \\ -\Delta_{p_2} u_2 = \lambda u_1^{\alpha_2} u_2^{\beta_2} + \delta h_2(x) \text{ in } \Omega \\ u_1, u_2 > 0 \text{ in } \Omega \\ u_1, u_2 = 0 \text{ on } \partial\Omega \end{cases},$$

where Ω is a bounded domain in $\mathbb{R}^N$ $(N \geqslant 2)$ having a smooth boundary $\partial\Omega$, $\lambda > 0$, $\delta \geqslant 0$ are parameters, and $h_i \in L^\infty(\Omega)$ is a nonnegative function. Here Δ_{p_i} stands for the p_i-Laplacian differential operator with $1 < p_i \leqslant N$. A solution of $(\mathcal{P})$ is understood in the weak sense, that is, a pair $(u_1, u_2) \in W_0^{1,p_1}(\Omega) \times W_0^{1,p_2}(\Omega)$, which are positive a.e. in Ω and satisfying

$$\int_\Omega |\nabla u_i|^{p_i-2} \nabla u_i \nabla \varphi_i \, dx = \int_\Omega (\lambda u_1^{\alpha_i} u_2^{\beta_i} + \delta h_i)\varphi_i \, dx, \text{ for all } \varphi_i \in W_0^{1,p_i}(\Omega), \ i = 1, 2.$$

We consider the system $(\mathcal{P})$ in a singular case assuming that

$$0 < \alpha_2 < p_1^* - 1, \ \ 0 < \beta_1 < p_2^* - 1 \text{ and } \ -1 < \alpha_1, \beta_2 < 0, \tag{1}$$

*hana.di@hotmail.fr
†brahim.khodja@univ-annaba.org
‡abdelkrim.moussaoui@univ-bejaia.dz https://orcid.org/0000-0003-1336-2257

where $p_i^* = \dfrac{Np_i}{N-p_i}$. This assumption makes system $(\mathcal{P})$ be cooperative, that is, for u_1(resp. u_2) fixed the right term in the first (resp. second) equation of $(\mathcal{P})$ is increasing in u_2 (resp. u_1).

Recently, singular cooperative system $(\mathcal{P})$ with $\delta = 0$ was mainly studied in [8, 9, 20]. In [20] existence and boundedness theorems for $(\mathcal{P})$ was established by using sub-supersolution method for systems combined with perturbation techniques. In [8] one gets existence, uniqueness, and regularity of a positive solution on the basis of an iterative scheme constructed through a sub-supersolution pair. In [9] an existence theorem involving sub-supersolution was obtained through a fixed point argument in a sub-supersolution setting. The semilinear case in $(\mathcal{P})$ (i.e. $p_i = 2$) was considered in [7, 13, 21] where the linearity of the principal part is essentially used. In this context, the singular system $(\mathcal{P})$ can be viewed as the elliptic counter-part of a class of Gierer-Meinhardt systems that models some biochemical processes (see, e.g. [21]). It can be also given an astrophysical meaning since it generalizes to the system the well-known Lane-Emden equation, where all exponents are negative (see [7]). For the one dimentional case ($N = 1$) we quote [15] and the references therein. The complementary situation for the system $(\mathcal{P})$ with respect to (1) is the so-called competitive system, which has recently attracted much interest. Relevant contributions regarding this topic can be found in [9, 18, 19]. For the regular case in $(\mathcal{P})$, that is when all the exponents are positive, we refer to [6, 22], while for quasilinear systems with singular weights we cite [2, 4] and their references.

It is worth pointing out that the aforementioned works have examined the subhomogeneous case $\Theta > 0$ of singular problem $(\mathcal{P})$ where

$$\Theta = (p_1 - 1 - \alpha_1)(p_2 - 1 - \beta_2) - \beta_1\alpha_2. \tag{2}$$

The constant Θ is related to system stability $(\mathcal{P})$ that behaves in a drastically different way, depending on the sign of Θ. For instance, for $\Theta < 0$ system $(\mathcal{P})$ is not stable in the sense that possible solutions cannot be obtained by iterative methods (see [5]).

Unlike the subhomogeneous case $\Theta > 0$ studied in the above references, the novelty of this paper is to establish the existence, regularity and nonexistence of (positive) solutions for singular problem $(\mathcal{P})$ by processing the two cases: 'homogeneous' when $\Theta = 0$ and 'superhomogeneous' if $\Theta < 0$. It should be noted that throughout this paper, $\Theta < 0$ (resp. $= 0$) means that $p_i - 1 - \alpha_i - \beta_i < 0$ (resp. $= 0$).

The existence result for problem $(\mathcal{P})$ is stated as follows.

Theorem 1. *Assume (1), $\Theta < 0$ (resp. $\Theta = 0$) and suppose that*

$$\inf_{\Omega} h_1(x), \ \inf_{\Omega} h_2(x) > 0. \tag{3}$$

Then, there is $\delta_0 > 0$ (resp. $\delta_0, \lambda_0 > 0$) such that, for all $\delta \in (0, \delta_0)$, problem $(\mathcal{P})$ possesses a (positive) solution (u_1, u_2) in $C_0^{1,\beta}(\overline{\Omega}) \times C_0^{1,\beta}(\overline{\Omega})$, for certain $\beta \in (0,1)$, verifying

$$u_i \geqslant cd(x) \quad \text{in } \Omega,$$

for some constant $c > 0$ and for all $\lambda > 0$ (resp. $\lambda \in (0, \lambda_0)$). Moreover, if $\Theta = \delta = 0$ and

$$\beta_1 = \frac{p_2}{p_1}(p_1 - 1 - \alpha_1) \quad \text{or} \quad \alpha_2 = \frac{p_1}{p_2}(p_2 - 1 - \beta_2), \tag{4}$$

then, there exists $\lambda_* > 0$ such that problem $(\mathcal{P})$ has no solution for every $\lambda \in (0, \lambda_*)$.

The main technical difficulty consists in the presence of singular terms in system $(\mathcal{P})$ with (1), expressed through (super-) homogeneous condition. Our approach is chiefly based on sub-supersolution method in its version for systems [3, section 5.5]. However, this method cannot

be directly implemented due to the presence of singular terms in $(\mathcal{P})$ under assumption (1). So, we first disturb system $(\mathcal{P})$ by introducing a parameter $\varepsilon > 0$. This gives rise to a regularized system for $(\mathcal{P})$ depending on ε whose study is relevant for our initial problem. By applying the sub-supersolution method, we show that the regularized system has a positive solution $(u_{1,\varepsilon}, u_{2,\varepsilon})$ in $C^{1,\beta}(\overline{\Omega}) \times C^{1,\beta}(\overline{\Omega})$ for some $\beta \in (0,1)$. It is worth noting that the choice of suitable functions with an adjustment of adequate constants is crucial in order to construct the sub-supersolution pair as well as to process the both cases $\Theta < 0$ and $\Theta = 0$. The (positive) solution (u_1, u_2) in $(W_0^{1,p_1}(\Omega) \cap L^\infty(\Omega)) \times (W_0^{1,p_2}(\Omega) \cap L^\infty(\Omega))$ of $(\mathcal{P})$ is obtained by passing to the limit as $\varepsilon \to 0$. This is based on a priori estimates, Fatou's Lemma and S_+-property of the negative p_i-Laplacian. The positivity of the solution (u_1, u_2) is achieved through assumption (3) while $C^{1,\beta}$-regularity is derived from the regularity result in [11].

The rest of the paper is organized as follows. Section 1 is devoted to the existence of solutions for the regularized system. Section 2 established the proof of the main result.

1. The regularized system

Given $1 < p < +\infty$, the space $L^p(\Omega)$ and $W_0^{1,p}(\Omega)$ are endowed with the usual norms $\|u\|_p = \Big(\int\limits_\Omega |u|^p \, dx\Big)^{1/p}$ and $\|u\|_{1,p} = \Big(\int\limits_\Omega |\nabla u|^p \, dx\Big)^{1/p}$, respectively. We will also use the space $C_0^{1,\beta}(\overline{\Omega}) = \big\{u \in C^{1,\beta}(\overline{\Omega}) : u = 0 \text{ on } \partial\Omega\big\}$ for a suitable $\beta \in (0,1)$.

In what follows, we denote by ϕ_{1,p_i} the positive eigenfunction associated with the principal eigenvalue λ_{1,p_i}, characterized by the minimum of Rayleigh quotient

$$\lambda_{1,p_i} = \inf_{u_i \in W_0^{1,p_i}(\Omega)\setminus\{0\}} \frac{\|\nabla u_i\|_{p_i}^{p_i}}{\|u_i\|_{p_i}^{p_i}}. \tag{5}$$

For a later use recall there exist constants $l_i, \hat{l}_i > 0$ such that

$$\hat{l}_1 \phi_{1,p_1}(x) \geqslant \phi_{1,p_2}(x) \geqslant \hat{l}_2 \phi_{1,p_1}(x) \text{ and } l_1 d(x) \geqslant \phi_{1,p_i}(x) \geqslant l_2 d(x) \text{ for all } x \in \Omega, \tag{6}$$

where $d(x) := dist(x, \partial\Omega)$ (see, e.g., [10]).

Let $\tilde{\Omega}$ be a bounded domain in $\mathbb{R}^N$ with a smooth boundary $\partial\tilde{\Omega}$ such that $\overline{\Omega} \subset \tilde{\Omega}$. Denote $\tilde{d}(x) := d(x, \partial\tilde{\Omega})$. By the definition of $\tilde{\Omega}$ there exists a constant $\rho > 0$ sufficiently small such that

$$\tilde{d}(x) > \rho \text{ in } \overline{\Omega}. \tag{7}$$

Define $w_i \in C^1(\overline{\tilde{\Omega}})$ the unique solution of the torsion problem

$$-\Delta_{p_i} w_i = 1 \text{ in } \tilde{\Omega}, \;\; w_i = 0 \text{ on } \partial\tilde{\Omega}, \tag{8}$$

satisfying the estimates

$$w_i(x) \geqslant c_0 \tilde{d}(x) \text{ in } \tilde{\Omega}, \tag{9}$$

for certain constant $c_0 \in (0,1)$ (see [12, Lemma 2.1]).

For a real constant $C > 1$, set

$$(\underline{u}_{i,\varepsilon}, \overline{u}_i) = (c_\varepsilon \phi_{1,p_i}, C^{-1} w_i), \;\; i = 1, 2, \tag{10}$$

where $c_\varepsilon > 0$ is a constant depending on $\varepsilon > 0$ such that

$$0 < c_\varepsilon < c_0 l_1^{-1} C^{-1}. \tag{11}$$

Then, by (10), (6) and (8), it is readily seen that

$$\overline{u}_i(x) = C^{-1}w_i(x) \geqslant C^{-1}c_0\tilde{d}(x) \geqslant C^{-1}c_0 d(x) \geqslant$$
$$\geqslant l_1^{-1}C^{-1}c_0\phi_{1,p_i}(x) \geqslant c_\varepsilon\phi_{1,p_i}(x) = \underline{u}_{i,\varepsilon}(x) \text{ in } \overline{\Omega}, \text{ for } i = 1, 2.$$

For every $\varepsilon \in (0, \varepsilon_0)$, with $\varepsilon_0 < 1$, let introduce the auxiliary problem

$$(\mathcal{P}_\varepsilon) \qquad \begin{cases} -\Delta_{p_1}u_1 = \lambda(u_1+\varepsilon)^{\alpha_1}(u_2+\varepsilon)^{\beta_1} + \delta h_1(x) & \text{in } \Omega \\ -\Delta_{p2}u_2 = \lambda(u_1+\varepsilon)^{\alpha_2}(u_2+\varepsilon)^{\beta_2} + \delta h_2(x) & \text{in } \Omega \\ u_1, u_2 = 0 & \text{on } \partial\Omega \end{cases},$$

which provides approximate solutions for the initial problem $(\mathcal{P})$.

Lemma 1. *Assume (1) and $h_1, h_2 \neq 0$ in Ω. Then, if $\Theta < 0$ (resp. $\Theta = 0$), there is a constant $\delta_0 > 0$ (resp. $\delta_0, \lambda_0 > 0$) such that for all $\delta \in (0, \delta_0)$, $(\overline{u}_1, \overline{u}_2)$ in (10) is a supersolution of $(\mathcal{P}_\varepsilon)$ for all $\lambda > 0$ (resp. $\lambda \in (0, \lambda_0)$) and all $\varepsilon \in (0, \varepsilon_0)$.*

Proof. Assume $\Theta < 0$ and set $\varepsilon_0 = C^{-1}$,

$$\delta_0 = \frac{1}{2}\min_{i=1,2}\left\{\frac{1}{C^{p_i-1}\left\|h_i\right\|_\infty}\right\}. \tag{12}$$

On account of (1), (7)–(10) and (12), for all $\delta \in (0, \delta_0)$ and $\varepsilon \in (0, \varepsilon_0)$, one derives

$$(\overline{u}_1+\varepsilon)^{-\alpha_1}(\overline{u}_2+\varepsilon)^{-\beta_1}(-\Delta_{p_1}\overline{u}_1 - \delta h_1) \geqslant \overline{u}_1^{-\alpha_1}(\overline{u}_2+\varepsilon_0)^{-\beta_1}(-\Delta_{p_1}\overline{u}_1 - \delta\left\|h_1\right\|_\infty) \geqslant$$
$$\geqslant C^{\alpha_1+\beta_1}(c_0\tilde{d}(x))^{-\alpha_1}(\left\|w_2\right\|_\infty + 1)^{-\beta_1}(C^{-(p_1-1)} - \delta_0\left\|h_1\right\|_\infty) \geqslant$$
$$\geqslant C^{\beta_1-(p_1-1-\alpha_1)}(c_0\rho)^{-\alpha_1}(\left\|w_2\right\|_\infty + 1)^{-\beta_1}(1 - \delta_0 C^{p_1-1}\left\|h_1\right\|_\infty) \geqslant$$
$$\geqslant \frac{1}{2}C^{\beta_1-(p_1-1-\alpha_1)}(c_0\rho)^{-\alpha_1}(\left\|w_2\right\|_\infty + 1)^{-\beta_1} \geqslant \lambda \text{ in } \overline{\Omega},$$

and similarly

$$(\overline{u}_1+\varepsilon)^{-\alpha_2}(\overline{u}_2+\varepsilon)^{-\beta_2}(-\Delta_{p_2}\overline{u}_2 - \delta h_2) \geqslant (\overline{u}_1+\varepsilon_0)^{-\alpha_2}\overline{u}_2^{-\beta_2}(-\Delta_{p_2}\overline{u}_2 - \delta\left\|h_2\right\|_\infty) \geqslant$$
$$\geqslant C^{\alpha_2+\beta_2}(\left\|w_1\right\|_\infty + 1)^{-\alpha_2}(c_0\tilde{d}(x))^{-\beta_2}(C^{-(p_2-1)} - \delta_0\left\|h_2\right\|_\infty) =$$
$$= C^{\alpha_2-(p_2-1-\beta_2)}(\left\|w_1\right\|_\infty + 1)^{-\alpha_2}(c_0\rho)^{-\beta_2}(1 - \delta_0 C^{p_2-1}\left\|h_2\right\|_\infty) \geqslant$$
$$\geqslant \tfrac{1}{2}C^{\alpha_2-(p_2-1-\beta_2)}(\left\|w_1\right\|_\infty + 1)^{-\alpha_2}(c_0\rho)^{-\beta_2} \geqslant \lambda \text{ in } \overline{\Omega},$$

for all $\lambda > 0$, provided $C > 1$ is sufficiently large. This shows that $(\overline{u}_1, \overline{u}_2)$ is a supersolution pair for problem $(\mathcal{P}_\varepsilon)$. If $\Theta = 0$, by repeating the argument above, the same conclusion can be drawn for $\lambda \in (0, \lambda_0)$ with a constant $\lambda_0 > 0$ that can be precisely estimated. This completes the proof. □

Lemma 2. *Assume (1) and $\Theta \leqslant 0$ hold. Then, $(\underline{u}_{1,\varepsilon}, \underline{u}_{2,\varepsilon})$ is a subsolution of $(\mathcal{P}_\varepsilon)$ for all $\lambda, \delta > 0$ and every $\varepsilon \in (0, \varepsilon_0)$.*

Proof. Fix $\varepsilon \in (0, \varepsilon_0)$. From (10) and (1), we obtain

$$\begin{aligned} &(\underline{u}_{1,\varepsilon}+\varepsilon)^{-\alpha_1}(\underline{u}_{2,\varepsilon}+\varepsilon)^{-\beta_1}(-\Delta_{p_1}\underline{u}_{1,\varepsilon} - \delta h_1) \leqslant \\ &\leqslant c_\varepsilon^{p_1-1}(c_\varepsilon\phi_{1,p_1}+\varepsilon_0)^{-\alpha_1}(c_\varepsilon\phi_{1,p_2}+\varepsilon)^{-\beta_1}\lambda_{1,p_1}\phi_{1,p_1}^{p_1-1} \leqslant \\ &\leqslant c_\varepsilon^{p_1-1}(\phi_{1,p_1}+\varepsilon_0)^{-\alpha_1}(c_\varepsilon\phi_{1,p_2}+\varepsilon)^{-\beta_1}\lambda_{1,p_1}\phi_{1,p_1}^{p-1} \leqslant \\ &\leqslant c_\varepsilon^{p_1-1}\varepsilon^{-\beta_1}(\left\|\phi_{1,p_1}\right\|_\infty + 1)^{-\alpha_1}\lambda_{1,p_1}\left\|\phi_{1,p_1}\right\|_\infty^{p_1-1} \leqslant \lambda \text{ in } \overline{\Omega} \end{aligned} \tag{13}$$

and similarly

$$\begin{aligned}
&(\underline{u}_{1,\varepsilon}+\varepsilon)^{-\alpha_2}(\underline{u}_{2,\varepsilon}+\varepsilon)^{-\beta_2}(-\Delta_{p_2}\underline{u}_{2,\varepsilon}-\delta h_2)\leqslant\\
&\leqslant(\underline{u}_{1,\varepsilon}+\varepsilon)^{-\alpha_2}(\underline{u}_{2,\varepsilon}+\varepsilon_0)^{-\beta_2}(\Delta_{p_2}\underline{u}_{2,\varepsilon})=\\
&=c_\varepsilon^{p_2-1}(c_\varepsilon\phi_{1,p_1}+\varepsilon)^{-\alpha_2}(\phi_{1,p_2}+\varepsilon_0)^{-\beta_2}\lambda_{1,p_2}\phi_{1,p_2}^{p_2-1}\leqslant\\
&\leqslant c_\varepsilon^{p_2-1}\varepsilon^{-\alpha_2}(\|\phi_{1,p_2}\|_\infty+\varepsilon_0)^{-\beta_2}\lambda_{1,p_2}\|\phi_{1,p_2}\|_\infty^{p_2-1}\leqslant\lambda\ \text{ in }\overline{\Omega},
\end{aligned}\tag{14}$$

provided $c_\varepsilon>0$ is sufficiently small. Gathering (13) and (14) together yields

$$-\Delta_{p_i}\underline{u}_{i,\varepsilon}\leqslant\lambda(\underline{u}_{1,\varepsilon}+\varepsilon)^{\alpha_i}(\underline{u}_{2,\varepsilon}+\varepsilon)^{\beta_i}+\delta h_i\ \text{ in }\overline{\Omega},$$

proving that $(\underline{u}_{1,\varepsilon},\underline{u}_{2,\varepsilon})$ in (10) is a subsolution pair for problem $(\mathcal{P}_\varepsilon)$. □

We state the following result regarding the regularized system.

Theorem 2. *Assume (1) and $h_1,h_2\neq 0$ in Ω. Then*

(a) *If $\Theta<0$ (resp. $\Theta=0$) there exist a constant $\delta_0>0$ (resp. $\delta_0,\lambda_0>0$) such that for all $\delta\in(0,\delta_0)$ system $(\mathcal{P}_\varepsilon)$ has a (positive) solution $(u_{1,\varepsilon},u_{2,\varepsilon})\in C_0^{1,\beta}(\overline{\Omega})\times C_0^{1,\beta}(\overline{\Omega})$, $\beta\in(0,1)$, satisfying*

$$u_{i,\varepsilon}(x)\leqslant\overline{u}_i(x)\text{ in }\Omega,\tag{15}$$

for all $\lambda>0$ (resp. $\lambda\in(0,\lambda_0)$), and every $\varepsilon\in(0,\varepsilon_0)$.

(b) *For $\Theta\leqslant 0$ and under assumption (3), if $\delta>0$, there exists a constant $c_0>0$, independent of ε, such that all solutions $(u_{1,\varepsilon},u_{2,\varepsilon})$ of system $(\mathcal{P}_\varepsilon)$ verify*

$$u_{i,\varepsilon}(x)\geqslant c_0d(x)\text{ for a.a. }x\in\Omega,\ \text{for all }\varepsilon\in(0,\varepsilon_0).\tag{16}$$

Proof. On the basis of Lemmas 1 and 2 together with [3, section 5.5] there exists a solution $(u_{1,\varepsilon},u_{2,\varepsilon})$ of problem $(\mathcal{P}_\varepsilon)$, for every $\varepsilon\in(0,\varepsilon_0)$. Moreover, applying the regularity theory (see [16]), we infer that $(u_{1,\varepsilon},u_{2,\varepsilon})\in C_0^{1,\beta}(\overline{\Omega})\times C_0^{1,\beta}(\overline{\Omega})$ for a suitable $\beta\in(0,1)$. This proves (a).

Now, according to (3), let $\sigma>0$ be a constant such that $\inf_\Omega h_1(x)$, $\inf_\Omega h_2(x)>\sigma$. Define z_i the only positive solution of

$$-\Delta_{p_i}z_i=\delta\sigma\ \text{ in }\Omega,\ z_i=0\ \text{ on }\partial\Omega,$$

which is known to satisfy $z_i(x)\geqslant c_2d(x)$ in Ω. Then it follows that $-\Delta_{p_i}u_\varepsilon\geqslant-\Delta_{p_i}z_i$ in Ω, $u_{i,\varepsilon}=z_i$ on $\partial\Omega$, for all $\varepsilon\in(0,\varepsilon_0)$, and therefore, the weak comparison principle ensures the assertion (b) holds true. □

2. Proof of the main result

Set $\varepsilon=\dfrac{1}{n}$ with any positive integer $n>1/\varepsilon_0$. From Theorem 2 with $\varepsilon=\dfrac{1}{n}$, there exists $u_{i,n}:=u_{i,\frac{1}{n}}$ such that

$$\langle-\Delta_{p_i}u_{i,n},\varphi_i\rangle=\lambda\int_\Omega\Big(u_{1,n}+\frac{1}{n}\Big)^{\alpha_i}\Big(u_{2,n}+\frac{1}{n}\Big)^{\beta_i}\varphi_i\,dx+\delta\int_\Omega h_i\varphi_i dx,\tag{17}$$

for all $\varphi_i \in W_0^{1,p_i}(\Omega), i = 1, 2$. Taking $\varphi_1 = u_{1,n}$ in (17), since $\alpha_1 < 0 < \beta_1$, we get

$$\begin{aligned}
&\|u_{1,n}\|_{1,p_1}^{p_1} = \lambda \int_\Omega \left(u_{1,n} + \frac{1}{n}\right)^{\alpha_1} \left(u_{2,n} + \frac{1}{n}\right)^{\beta_1} u_{1,n}\, dx + \int_\Omega \delta h_1 u_{1,n}\, dx \leqslant \\
&\leqslant \lambda \int_\Omega u_{1,n}^{\alpha_1+1} \left(u_{2,n} + 1\right)^{\beta_1}\, dx + \delta \|h_1\|_\infty \int_\Omega u_{1,n}\, dx \leqslant \\
&\leqslant \lambda \int_\Omega \overline{u}_1^{\alpha_1+1} \left(\overline{u}_2 + 1\right)^{\beta_1}\, dx + \delta \|h_1\|_\infty \int_\Omega \overline{u}_1\, dx \leqslant \\
&\leqslant \lambda |\Omega| (\|\overline{u}_1\|_\infty^{\alpha_1+1} \left(\|\overline{u}_2\|_\infty + 1\right)^{\beta_1} + \delta \|h_1\|_\infty \|\overline{u}_1\|_\infty).
\end{aligned} \tag{18}$$

Hence, $\{u_{1,n}\}$ is bounded in $W_0^{1,p_1}(\Omega)$. Similarly, we derive that $\{u_{2,n}\}$ is bounded in $W_0^{1,p_2}(\Omega)$. We are thus allowed to extract subsequences (still denoted by $\{u_{i,n}\}$) such that

$$u_{i,n} \rightharpoonup u_i \text{ in } W_0^{1,p_i}(\Omega),\ i = 1, 2. \tag{19}$$

The convergence in (19) combined with Rellich embedding Theorem and (15)–(16) entails

$$c_0 d(x) \leqslant u_i(x) \leqslant \overline{u}_i(x) \text{ in } \Omega. \tag{20}$$

Inserting $\varphi_i = u_{i,n} - u_i$ in (17) yields

$$\langle -\Delta_{p_i} u_{i,n}, u_{i,n} - u_i \rangle = \int_\Omega \left[\lambda \Big(u_{1,n} + \frac{1}{n}\Big)^{\alpha_i} \Big(u_{2,n} + \frac{1}{n}\Big)^{\beta_i} + \delta h_i\right](u_{i,n} - u_i)\, dx.$$

We claim that

$$\lim_{n\to\infty} \langle -\Delta_{p_i} u_{i,n}, u_{i,n} - u_i \rangle \leqslant 0.$$

Indeed, from (15), (16) and (10), we have

$$\begin{aligned}
&\left|\left(\Big(u_{1,n} + \frac{1}{n}\Big)^{\alpha_1} \Big(u_{2,n} + \frac{1}{n}\Big)^{\beta_1} + \delta h_1\right)(u_{1,n} - u_1)\right| \leqslant \\
&\leqslant \big(u_{1,n}^{\alpha_1} \left(u_{2,n} + 1\right)^{\beta_1} + \delta h_1\big)(|u_{1,n}| + |u_1|) \leqslant \\
&\leqslant 2\big((c_0 d(x))^{\alpha_1} \left(\overline{u}_2 + 1\right)^{\beta_1} + \delta h_1\big)\overline{u}_1 \leqslant \\
&\leqslant 2\big((c_0 d(x))^{\alpha_1} \left(\|\overline{u}_2\|_\infty + 1\right)^{\beta_1} + \delta \|h_1\|_\infty\big) \|\overline{u}_1\|_\infty \leqslant \hat{C}_0 d(x)^{\alpha_1} \text{ in } \Omega,
\end{aligned}$$

with some positive constant $\hat{C}_0$. Then, (1) together with Lemma in [14, page 726] imply that

$$\left(\Big(u_{1,n} + \frac{1}{n}\Big)^{\alpha_1} \Big(u_{2,n} + \frac{1}{n}\Big)^{\beta_1} + \delta h_1\right)(u_{1,n} - u_1) \in L^1(\Omega). \tag{21}$$

Using (19), (21) and applying Fatou's Lemma, it follows that

$$\begin{aligned}
&\limsup_{n\to\infty} \int_\Omega \left(\left(u_{1,n} + \frac{1}{n}\right)^{\alpha_1} \left(u_{2,n} + \frac{1}{n}\right)^{\beta_1} + \delta h_1\right)(u_{1,n} - u_1)\, dx \leqslant \\
&\leqslant \int_\Omega \limsup_{n\to\infty} \left(\left(\left(u_{1,n} + \frac{1}{n}\right)^{\alpha_1} \left(u_{2,n} + \frac{1}{n}\right)^{\beta_1} + \delta h_1\right)(u_{1,n} - u_1)\right) dx \to 0,
\end{aligned}$$

showing that $\limsup_{n\to\infty} \langle -\Delta_{p_1} u_{1,n}, u_{1,n} - u_1 \rangle \leqslant 0$. Likewise, we prove that

$$\limsup_{n\to\infty} \langle -\Delta_{p_2} u_{2,n}, u_{2,n} - u_2 \rangle \leqslant 0.$$

Then the S_+-property of $-\Delta_{p_i}$ on $W_0^{1,p_i}(\Omega)$ (see, e.g., [17, Proposition 3.5]) guarantees that

$$u_{i,n} \longrightarrow u_i \ \text{ in } W_0^{1,p_i}(\Omega),\ i=1,2. \tag{22}$$

On account of (17), besides (22), the next step is to verify that

$$\lim_{n\to\infty}\int_\Omega \Big(u_{1,n}+\frac{1}{n}\Big)^{\alpha_i}\ \Big(u_{2,n}+\frac{1}{n}\Big)^{\beta_i}\varphi_i\ dx=\int_\Omega u_1^{\alpha_i}u_2^{\beta_i}\varphi_i\ dx, \tag{23}$$

for all $\varphi_i\in W_0^{1,p_i}(\Omega)$. By (15), (16), (1) and (20), it holds

$$\left|\Big(u_{1,n}+\frac{1}{n}\Big)^{\alpha_1}\ \Big(u_{2,n}+\frac{1}{n}\Big)^{\beta_1}\varphi_1\right|\leqslant (c_0d(x))^{\alpha_1}(\|\overline{u}_2\|_\infty+1)^{\beta_1}|\varphi_1|$$

and

$$\left|\Big(u_{1,n}+\frac{1}{n}\Big)^{\alpha_2}\ \Big(u_{2,n}+\frac{1}{n}\Big)^{\beta_2}\varphi_2\right|\leqslant (\|\overline{u}_1\|_\infty+1)^{\alpha_2}(c_0d(x))^{\beta_2}|\varphi_2|.$$

Then, by (1) together with Hardy-Sobolev inequality (see, e.g., [1, Lemma 2.3]), assertion (23) stem from Lebesgue's dominated convergence Theorem. Hence we may pass to the limit in (17) to conclude that (u_1,u_2) is a solution of problem $(\mathcal{P})$ satsifying (20). Furthermore, using (1), (20) and (10), one has

$$\begin{aligned}&u_1^{\alpha_1}u_2^{\beta_1}+\delta h_1\leqslant u_1^{\alpha_1}\overline{u}_2^{\beta_1}+\delta\,\|h_1\|_\infty\leqslant\\&\leqslant (c_0d(x))^{\alpha_1}\,\|\overline{v}\|_\infty^{\beta_1}+\delta\,\|h_1\|_\infty\,d(x)^{\alpha_1-\alpha_1}\leqslant\\&\leqslant C_1'd(x)^{\alpha_1}\ \text{for all } x\in\Omega\end{aligned} \tag{24}$$

and

$$\begin{aligned}&u_1^{\alpha_2}u_2^{\beta_2}+\delta h_2\leqslant \overline{u}_1^{\alpha_2}u_2^{\beta_2}+\delta\,\|h_2\|_\infty\leqslant\\&\leqslant \|\overline{u}_1\|_\infty^{\alpha_2}\,(c_0d(x))^{\beta_2}+\delta\,\|h_2\|_\infty\,d(x)^{\beta_2-\beta_2}\leqslant\\&\leqslant C_2'd(x)^{\beta_2}\ \text{for all } x\in\Omega,\end{aligned} \tag{25}$$

for certain positive constants C_1' and C_2'. Hence, (1) enable us to apply Lemma 3.1 in [11] to infer that $(u,v)\in C_0^{1,\beta}(\overline{\Omega})\times C_0^{1,\beta}(\overline{\Omega})$ for some $\beta\in(0,1)$.

We are left with the task of determining the nonexistence result stated in Theorem 1. Arguing by contradiction and assume that (u_1,u_2) is a positive solution of problem $(\mathcal{P})$ with $\delta=0$. Multiplying in $(\mathcal{P})$ by u_i, integrating over Ω, applying Young inequality with $\alpha_1,\beta_2>-1$, we get

$$\int_\Omega|\nabla u_1|^{p_1}\,dx=\lambda\int_\Omega u_1^{\alpha_1+1}u_2^{\beta_1}dx\leqslant\lambda\int_\Omega\left(\frac{\alpha_1+1}{p_1}u_1^{p_1}+\frac{p_1-1-\alpha_1}{p_1}u_2^{\frac{\beta_1p_1}{p_1-1-\alpha_1}}\right)dx \tag{26}$$

and

$$\int_\Omega|\nabla u_2|^{p_2}\,dx=\lambda\int_\Omega u_1^{\alpha_2}u_2^{\beta_2+1}dx\leqslant\lambda\int_\Omega\left(\frac{p_2-1-\beta_2}{p_2}u_1^{\frac{\alpha_2p_2}{p_2-1-\beta_2}}+\frac{\beta_2+1}{p_2}u_2^{p_2}\right)dx. \tag{27}$$

Adding (26) with (27), according to (4), this is equivalent to

$$\|\nabla u_1\|_{p_1}^{p_1}+\|\nabla u_2\|_{p_2}^{p_2}\leqslant\lambda\left[\Big(\frac{\alpha_1+1}{p_1}+\frac{p_2-1-\beta_2}{p_2}\Big)\,\|u_1\|_{p_1}^{p_1}+\Big(\frac{\beta_2+1}{p_2}+\frac{p_1-1-\alpha_1}{p_1}\Big)\,\|v\|_{p_2}^{p_2}\right]. \tag{28}$$

Since $\Theta=0$, observe from (4) that

$$\begin{cases}\dfrac{\alpha_1+1}{p_1}+\dfrac{p_2-1-\beta_2}{p_2}=\dfrac{\alpha_1+\alpha_2+1}{p_1}\\[2ex]\dfrac{\beta_2+1}{p_2}+\dfrac{p_1-1-\alpha_1}{p_1}=\dfrac{\beta_1+\beta_2+1}{p_2}.\end{cases} \tag{29}$$

Then gathering (5), (28) and (29) together yields

$$\left(\lambda_{1,p_1} - \frac{\alpha_1+\alpha_2+1}{p_1}\lambda\right)\|u_1\|_{p_1}^{p_1} + \left(\lambda_{1,p_2} - \frac{\beta_1+\beta_2+1}{p_2}\lambda\right)\|u_2\|_{p_2}^{p_2} \leqslant 0$$

which is a contradiction for

$$\lambda < \lambda_* = \min\left\{\frac{p_1}{\alpha_1+\alpha_2+1}\lambda_{1,p_1}, \frac{p_2}{\beta_1+\beta_2+1}\lambda_{1,p_2}\right\}.$$

Thus, problem $(\mathcal{P})$ has no solution for $\lambda < \lambda_*$, which completes the proof.

References

[1] C.O.Alves, F.J.S.A.Corrêa, On the existence of positive solution for a class of singular systems involving quasilinear operators, *Appl. Math. Comput.*, **185**(2007), 727–736.

[2] S.Boulaaras, R.Guefaifia, T.Bouali, Existence of positive solutions for a class of quasilinear singular ellipticsystems involving Caffarelli-Kohn-Nirenberg exponent with sign-changing weight functions, *Indian J. Pure Appl. Math.*, **49**(2018), no. 4, 705–715.

[3] S.Carl, V.K.Le, D.Motreanu, Nonsmooth variational problems and their inequalities. Comparaison principles and applications, Springer, New York, 2007.

[4] Z.Deng, R.Zhang, Y.Huang, Multiple symmetric results for singular quasilinear elliptic systems with critical homogeneous nonlinearity, *Math. Meth. App. Sci.*, **40**(2017), no. 5, 1538–1552.

[5] P.Clément, J.Fleckinger, E.Mitidieri, F.de Thelin, Existence of Positive Solutions for a Non-variational Quasilinear Elliptic System, *J. Diff. Eqts.*, **166**(2000), 455–477.

[6] C.Chen, On positive weak solutions for a class of quasilinear elliptic systems, *Nonl. Anal.*, **62**(2005), no. 4, 751–756.

[7] M.Ghergu, Lane-Emden systems with negative exponents, *J. Funct. Anal.*, **258**(2010), 3295–3318.

[8] J.Giacomoni, J.Hernandez, A.Moussaoui, Quasilinear and singular systems: the cooperative case, *Contemporary Math.*, **540**(2011), 79–94.

[9] J.Giacomoni, J.Hernandez, P.Sauvy, Quasilinear and singular elliptic systems, *Advances Nonl. Anal.*, **2**(2013), 1–41.

[10] J.Giacomoni, I.Schindler, P.Takac, Sobolev versus H ölder local minimizers and existence of multiple solutions for a singular quasilinear equation, *A. Sc. N. Sup. Pisa*, **6**(2007), no. 5, 117–158.

[11] D.D.Hai, On a class of singular p-Laplacian boundary value problems, *J. Math. Anal. Appl.*, **383**(2011), 619–626.

[12] D.D.Hai, H.Wang, Nontrivial solutions for p-Laplacian systems, *J. Math. Anal. Appl.*, **330**(2007), 186–194.

[13] J.Hernandez, F.J.Mancebo, J.M.Vega, Positive solutions for singular semilinear elliptic systems, *Adv. Diff. Eqts.*, **13**(2008), 857–880

[14] A.C.Lazer, P.J.Mckenna, On a singular nonlinear elliptic boundary-value problem, *Proc. American Math. Soc.*, **111**(1991), no. 3, 721–730.

[15] Y-H.Lee, X.Xu, Global Existence Structure of Parameters for Positive Solutions of a Singular (p_1, p_2)-Laplacian System, *Bull. Malays. Math. Sci. Soc.*, **42**(2019), no. 3, 1143–1159.

[16] G.M.Lieberman, Boundary regularity for solutions of degenerate elliptic equations, *Nonlinear Anal.*, **12**(1988), 1203–1219.

[17] D.Motreanu, V.V.Motreanu, N.S.Papageorgiou, Multiple constant sign and nodal solutions for Nonlinear Neumann eigenvalue problems, *Ann. Sc. Norm. Super. Pisa Cl. Sci.*, **10**(2011), no. 5, 729–755.

[18] D.Motreanu, A.Moussaoui, A quasilinear singular elliptic system without cooperative structure, *Act. Math. Sci.*, **34**(2014), no. 3, 905–916.

[19] D.Motreanu, A.Moussaoui, An existence result for a class of quasilinear singular competitive elliptic systems, *Applied Math. Letters*, **38**(2014), 33–37.

[20] D.Motreanu, A.Moussaoui, *Complex Variables and Elliptic Equations*, **59**(2014), 285–296. DOI: 10.1080/17476933.2012.744404

[21] A.Moussaoui, B.Khodja, S.Tas, A singular Gierer-Meinhardt system of elliptic equations in R^N, *Nonlinear Anal.*, **71**(2009), 708–716.

[22] R.S.Rodrigues, Positive solutions for classes of positone/semipositone systems with multiparameters, *Electronic Journal of Differential Equations*, (2013), no. 192, 1–10.

Сингулярные квазилинейные эллиптические системы с (супер-)однородным условием

Хана Диди
Брахим Ход
Университет Баджи-Мохтар Аннаба
Аннаба, Алжир
Абделькрим Муссауи
Университет Мира Беджайа
Беджайа, Алжир

Аннотация. В данной работе мы устанавливаем существование (несуществование) и регулярность положительных решений для класса сингулярных квазилинейных эллиптических систем, подчиняющихся (супер-)однородному условию. Подход основан на методах субсуперрешений для систем квазилинейных сингулярных уравнений в сочетании с аргументами возмущения, включающими сингулярные члены.

Ключевые слова: сингулярная система, p-лапласиан, субсуперрешение, регулярность.

DOI: 10.17516/1997-1397-2020-13-2-160-169
УДК 519.873

A Recursive Algorithm for Estimating the Correlation Matrix of the Interference Based on the QR Decomposition

Valery N. Tyapkin*
Dmitry D. Dmitriev†
Andrey B. Gladyshev‡
Peter Yu. Zverev§
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 02.11.2019, received in revised form 10.12.2019, accepted 20.01.2020

Abstract. Many tasks of digital signal processing require the implementation of matrix operations in real time. These are operations of matrix inversion or solving systems of linear algebraic or differential equations (Kalman filter). The transition to the implementation of digital signal processing on programmable logic device (FPGAs), as a rule, involves calculations based on the representation of numbers with a fixed point. This makes solving spatio-temporal processing problems practically impossible based on conventional computational methods. The article discusses the implementation of spatial-temporal signal processing algorithms in satellite broadband systems using QR decomposition. The technologies of CORDIC computations required for recurrent QR decomposition when used together in systolic algorithms are presented.

Keywords: phased antenna array, adaptive algorithms, Kalman filter, recursive least squares algorithm (RLS), QR decomposition, systolic algorithm.

Citation: V.N.Tyapkin, D.D.Dmitriev, A.B.Gladyshev, P.Yu.Zverev, A Recursive Algorithm for Estimating the Correlation Matrix of the Interference Based on the QR Decomposition, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 160–169. DOI: 10.17516/1997-1397-2020-13-2-160-169.

Introduction

Modern satellite broadband radio communication systems have a significant drawback - low noise immunity. The solution to this problem is based on the use of phased array antennas. Such antennas are controlled using adaptive algorithms, the parameters of which can be quickly changed in accordance with the emerging interference environment.

To work effectively in real conditions of parametric a priori uncertainty and a dynamic change in the statistical characteristics of interference, adaptive systems are required, the parameters of which can be quickly changed in accordance with the interference environment. Currently, theoretically substantiated and tested in practice, the methods of the Markov theory of optimal filtration. This theory has been fully and strictly developed in a number of books and articles

*tyapkin58@mail.ru
†dmitriev121074@mail.ru https://orcid.org/0000-0001-6438-6094
‡a-glonass@yandex.ru
§peter676@mail.ru

[1–12]. In [6, 10, 11], the solution to the problem of adaptive filtering of signals based on the use of the lemma on the inversion of the correlation matrix (MIL) of input signals of an adaptive filter is considered. This solution leads to a recursive least squares (RLS) algorithm.

The same problem can also be solved by reducing the matrix of input samples of the adaptive filter signals to a triangular form. In this case, the range of numbers involved in the calculations is reduced by comparing the solution to this problem with the estimation of the inverse correlation matrix of interference using MIL. This increases the stability of QRD RLS algorithms.

Assume that the signal and interference affect the input of a multichannel M-element phased array antenna. The set of signals from the outputs of the M-element antenna array is described by the time function $y_1(t), y_2(t), \ldots, y_m(t)$ and form a column vector $y_1(t) = = [y_1(t), y_2(t), \ldots, y_m(t)]^T$. Moreover, a single-channel reception ($M = 1$) is considered as a special case of multi-channel.

Discretization of the received useful and interference signals is performed at the radio frequency f_0. A feature of this is the small sampling interval Δt, which is approximately half the period of the carrier frequency $T_d = 1/f_0$, $\Delta t \approx 1/2f_0 = T_0/2$. Discrete interference values obtained from the antenna are random numbers that are conveniently represented as a column vector $y_i = [y_i(k\Delta t)] = [y_i(k)]$, $k = \overline{1, L}$, where L determines the duration of the observation interval $T - L = T/\Delta t$. In the case of multichannel reception, the vector of received oscillations will have the following form $\mathbf{Y} = (\mathbf{Y}_1\mathbf{Y}_2 \ldots \mathbf{Y}_k \ldots \mathbf{Y}_T)^T$ and dimension $(T \times M)$.

Most QR decomposition algorithms are based on Householder reflection and Givens rotations [13]. For the implementation of space-time processing, the most useful is the recursive version of the Givens method, which provides updates to the solutions at the rate of arrival of the input samples of the signal. High real-time performance provides a systolic version of the QR algorithm using pipelined implementation of Givens rotations on FPGA. High speed fixed-point number calculations on FPGAs are provided by the CORDIC processor. The principle of its operation differs significantly from the arithmetic-logical devices of existing processors. To implement Givens rotation, 10 shift-addition operations are sufficient. In this case, an accuracy sufficient to achieve an interference suppression ratio of more than 50 dB is ensured.

Thus, the development of this research area promises a significant improvement in the quality of reception and processing of broadband signals and noise immunity based on the existing element base and is relevant. Consider the implementation of spatial-temporal signal processing algorithms using QR decomposition. To solve the problem of recurrent QR decomposition, we will develop CORDIC computing technology in systolic algorithms.

1. Recursive adaptation algorithm using QR decomposition

A recursive adaptation algorithm using QR decomposition estimates the filter coefficients at the current time step through the calculated filter coefficient at the previous step. Due to its recursive nature, the algorithm is called QRD — a recursive QR decomposition algorithm. QR factorization consists in reducing a linear system to a triangular one. For this, the original matrix is represented as the product of the upper triangular matrix $\mathbf{R}$ and the orthogonal matrix $\mathbf{Q}$.

Consider a system of linear equations

$$\mathbf{A}_\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\mathbf{A} - (n \times m)$ is the matrix, $\mathbf{x}$ is the vector of derivatives (for example, weights for the adaptive antenna array), $\mathbf{b}$ is the m-vector. The QR decomposition of matrix $\mathbf{A}$ of size $(n \times m)$

for any $n \gg m$ can be described as:

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{R}, \tag{2}$$

$\mathbf{Q}$ is a unitary matrix: $\mathbf{Q} \cdot \mathbf{Q^H} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, $\mathbf{R}$ is $(n \times m)$ the upper right triangular matrix. Equation (2) can be written in divided form:

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{R} = \mathbf{Q} \begin{bmatrix} \mathbf{R_1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{Q_1} & \mathbf{Q_2} \end{bmatrix} \begin{bmatrix} \mathbf{R_1} \\ \mathbf{0} \end{bmatrix} = \mathbf{Q_1} \cdot \mathbf{R_1}, \tag{3}$$

where $\mathbf{R}_1$ is the $(m \times m)$ triangular matrix, $\mathbf{Q}_1$ is the $(m \times n)$ matrix and $\mathbf{Q}_2$ is the $((n-m) \times m)$ matrix.

In the spatio-temporal processing of broadband signals, the system of equations is usually redefined because $n \gg m$. Solution (2) minimizing the norm of the residual $\|\mathbf{Ax} - \mathbf{b}\|$ has the form $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.

Substitution of equality $\mathbf{A} = \mathbf{Q}^T\mathbf{R}$ gives the following form

$$\mathbf{A} = (\mathbf{R}^T\mathbf{Q}^T\mathbf{QR})^{-1}\mathbf{R}^T\mathbf{Q}^T\mathbf{b} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{b}.$$

Here the triangular $(m \times m)$-matrix is subject to circulation, which requires $(m^2)/2$ operations of addition and multiplication. Let us synthesize a recursive algorithm for estimating the correlation matrix of interference based on QR decomposition From the theory of matrices it is known that there exists a unitary matrix $\mathbf{Q}_k(k)$, that for any $\mathbf{A}_{kM}$ matrix can be obtained by decomposition

$$\mathbf{Q}_k^H \mathbf{A}_{kM} = \begin{bmatrix} \mathbf{F}_M \\ \mathbf{0}_{(T-M)M} \end{bmatrix}, \tag{4}$$

where $\mathbf{F}_M$ is the upper triangular square matrix, called the Cholesky decomposition, $\mathbf{0}_{(T-M)M}$ is the zero rectangular matrix. The superscript H means complex conjugation and transposition.

For the unitary matrices $\mathbf{Q}_k(k)$, the equalities $\mathbf{Q}_k(k) \cdot \mathbf{Q}_k^H(k) = \mathbf{Q}_k^H(k) \cdot \mathbf{Q}_k(k) = \mathbf{I}$ and $\mathbf{Q}_k^H(k) = \mathbf{Q}_k^{-1}(k)$.

If designated

$$\mathbf{A}_{kM} = \mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k), \tag{5}$$

where

$$\mathbf{\Lambda}_k^{0.5} = diag\left\{\sqrt{\lambda^{k-1}}, \sqrt{\lambda^{k-2}}, \cdots \sqrt{\lambda^1}, 1\right\} = \begin{bmatrix} \sqrt{\lambda^{k-1}} & 0 & \cdots & 0 & 0 \\ 0 & \sqrt{\lambda^{k-2}} & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda} & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{6}$$

The parameter λ is used to weight the signals and allows you to take into account changes in the statistics of signals if they are non-stationary and their statistical parameters change over time. The parameter λ is also called the exponential weighting parameter or the "orgetting factor" parameter. Its value is usually limited by the limits $(1 - 0.4K) \leqslant \lambda \leqslant 1$ [14, 15]. For example, at $\lambda = 0.9$, $\lambda_0 = 1$, $\lambda_1 = 0.9$, $\lambda_2 = 0.81$, $\lambda_3 = 0.72$, $\lambda_4 = 0.66$, $\lambda_5 = 0.59, \ldots, \lambda_{10} = 0.35, \ldots, \lambda_{20} = 0.12, \ldots, \lambda_{30} = 0.04, \ldots, \lambda_{40} = 0.015, \ldots, \lambda_{50} = 0.005, \ldots, \lambda_{100} = 0.000027$, that is, the parameter λ determines the "memory" of the algorithm for solving the recursive problem.

For $0 < \lambda < 1$ and small values of the difference $k - i$, the summed terms are weighted with large weights, and for large values of this difference, with smaller weights. For $\lambda = 1$ this

"memory" is equal to k samples of the observed signals. For $0 < \lambda < 1$, the contribution of the same i-th samples to the sum is different for different λ. This contribution is greater for large λ and smaller for smaller λ. That is, with decreasing λ, the effective memory decreases and vice versa. Substituting equation (5) in (4) we obtain

$$\mathbf{Q}_k^H \mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k) = \begin{bmatrix} \hat{\mathbf{R}}_M(k) \\ \mathbf{0}_{(T-M)M} \end{bmatrix}, \tag{7}$$

where

$$\mathbf{Y}_{Mk}(k)\mathbf{\Lambda}_k^{0.5}(k)_{kM}\mathbf{Q}_k(k) = \begin{bmatrix} \hat{\mathbf{R}}_M(k) \\ \mathbf{0}_{(T-M)M} \end{bmatrix}^{\mathrm{H}} = \left[\hat{\mathbf{R}}_M^{\mathrm{H}}(k), \mathbf{0}_{(T-M)M}^{\mathrm{T}}\right]. \tag{8}$$

Reducing the matrix $\hat{\mathbf{R}}_M(k)$ to a triangular form using the observation matrix $\mathbf{Y}_{Mk}^H(k)$ can be performed in various ways, the main of which is Givens rotation. The matrix $\hat{\mathbf{R}}_M(k)$ can be obtained recursively in time, performing calculations for the k-th iteration using the results from the previous, $(k-1)$-th iteration. This is as follows.

Let us assume that at iteration k-1 there is a decomposition:

$$\mathbf{Q}_{k-1}^H(k-1)\mathbf{\Lambda}_{k-1}^{0.5}(k-1)\mathbf{Y}_{M(k-1)}^H(k-1) = \begin{bmatrix} \hat{\mathbf{R}}_M \ (k-1) \\ \mathbf{0}_{(k-1-M)M} \end{bmatrix} \tag{9}$$

and conversion is required

$$\mathbf{Q}_k^H(k)\mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k) = \begin{bmatrix} \hat{\mathbf{R}}_M \ (k) \\ \mathbf{0}_{(k-M)M} \end{bmatrix}. \tag{10}$$

Using the result of (9), we define the matrix

$$\tilde{\mathbf{Q}}_k^H(k) = \begin{bmatrix} \mathbf{Q}_{k-1}^H(k-1) & \mathbf{0}_{k-1} \\ \mathbf{0}_{k-1}^T & 1 \end{bmatrix}. \tag{11}$$

If the matrix $\mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k)$ multiplied from left on the matrix (11), this operation modifies equation (9), adding it to the matrix in the right side of $(k+1)-w$ (bottom) row:

$$\begin{aligned} &\tilde{\mathbf{Q}}_k^H(k)\mathbf{\Lambda}_{k-1}^{0.5}(k)\mathbf{Y}_{Mk}^H(k) = \tilde{\mathbf{Q}}_k^H(k)\begin{bmatrix} \lambda^{0.5}\mathbf{\Lambda}_{k-1}^{0.5}(k-1)\mathbf{Y}_{M(k-1)}^H(k-1) \\ \mathbf{Y}_M^H(k) \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{Q}_{k-1}^H\lambda^{0.5}\mathbf{\Lambda}_{k-1}^{0.5}(k-1)\mathbf{Y}_{M(k-1)}^H(k-1) \\ \mathbf{Y}_M^H(k) \end{bmatrix} = \begin{bmatrix} \lambda^{0.5}\hat{\mathbf{R}}_M(k-1) \\ \mathbf{0}_{(k-1-M)M} \\ \mathbf{Y}_M^H(k) \end{bmatrix}. \end{aligned} \tag{12}$$

To perform the transformation (10), in equation (12) it is necessary to zero the last line. From equations (10) and (12) it follows that

$$\begin{aligned} &\mathbf{Q}_k^H\mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k) = \hat{\mathbf{Q}}_k^H(k)\tilde{\mathbf{Q}}_k^H(k)\mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k) = \\ &= \hat{\mathbf{Q}}_k^H(k) = \begin{bmatrix} \lambda^{0.5}\hat{\mathbf{R}}_M(k-1) \\ \mathbf{0}_{(k-1-M)M} \\ \mathbf{Y}_M^H(k) \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{R}}_M(k) \\ \mathbf{0}_{(k-1-M)M} \\ \mathbf{0}_M \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{R}}_M(k) \\ \mathbf{0}_{(k-M)M} \end{bmatrix}, \end{aligned} \tag{13}$$

where the matrix $\mathbf{Q}_k^H(k)$ is the product of two matrices

$$\mathbf{Q}_k^H(k) = \hat{\mathbf{Q}}_k^H(k)\tilde{\mathbf{Q}}_k^H(k). \tag{14}$$

Thus, the reduction of matrix $\mathbf{\Lambda}_k^{0.5}(k)\mathbf{Y}_{Mk}^H(k)$ to a triangular form using the matrix $\mathbf{Q}_k^H(k)$ (10) at iteration k can be done by zeroing the last row in the matrix (12) using the result of reducing the matrix $\mathbf{\Lambda}_{k-1}^{0.5}(k-1)\mathbf{Y}_{M(k-1)}^H(k-1)$ to a triangular form obtained at iteration $k-1$. This zeroing is carried out using matrix $\hat{\mathbf{Q}}_k^H(k)$, which is a product of matrices composed of Givens rotation matrices.

The recursive relationship between $\hat{\mathbf{R}}_M(k-1)$ and $\hat{\mathbf{R}}_M(k)$in a more compact form, i.e., when k matrices with a fixed number of elements $(M+1)\times(M+1)$ are used at each iteration, can be represented using the following equation

$$\mathbf{G}_{M+1}(k)\begin{bmatrix}\lambda^{0.5}\hat{\mathbf{R}}(k-1)\\ \mathbf{Y}_M^{\mathrm{H}}(k)\end{bmatrix}=\begin{bmatrix}\hat{\mathbf{R}}(k)\\ \mathbf{0}_M^{\mathrm{T}}\end{bmatrix}, \tag{15}$$

where matrix $\mathbf{G}_{M+1}(k)$ is unitary. This matrix can be formed using Givens spins. The structure of matrices $\mathbf{G}_{M+1}(k)$ is a "compressed" to size $(M+1)\times(M+1)$ matrix $\hat{\mathbf{Q}}_k^H(k)$ with a variable number of elements $(k)\times(k)$.

The elements of Givens matrices are determined from equation (15), where

$$\mathbf{G}_{M+1}(k)=\prod_{i=1}^{M}\mathbf{G}_{M+1}^i(k), \tag{16}$$

Givens $\mathbf{G}_M^i(k)$ transformation is determined by the plane rotation matrices of the form:

$$\mathbf{G}(i,j)=\begin{bmatrix}1 & & 0 & & & & & & \\ & \ddots & & & & & & 0 & \\ 0 & & 1 & & & & & & \\ & & & c & \dots & s & & & \\ & & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & & & & 1 & & \\ & & & -s & \dots & c & & & \\ & & & & & & & 1 & \\ & 0 & & & & & & \ddots & \\ & & & & & & & & 1\end{bmatrix}. \tag{17}$$

Matrix $\mathbf{G}_{i,j}$ with fixed values of $i,j\in\{1,2,\dots,m-1\}$ differs from the identity n–matrix $\mathbf{E}$ in that in it the 2×2-submatrix $\tilde{\mathbf{E}}$ occupying the cell formed by the intersection of the i-th and j-th rows and columns is replaced by the submatrix $\tilde{\mathbf{G}}_i=\begin{pmatrix}c & -s^*\\ s & c\end{pmatrix}$, with elements c and s satisfying the condition

$$s^2+c^2=1. \tag{18}$$

With this normalization condition, matrix $\tilde{\mathbf{G}}_i$ and matrix $\mathbf{G}_i$ are orthogonal. The elements c and s can be interpreted as the cosine and sine of a certain angle of rotation transformation.

Using a sequence of such orthogonal transformations, matrices $\mathbf{G}_1,\mathbf{G}_2,\dots,\mathbf{G}_{m-1}$ of the form (17) can be reduced to the right triangular form by sequentially canceling the subdiagonal elements in the first, second, . . . , $(n-1)$-st columns. We consider the first of N steps leading to

transformation (15), for which matrix $\mathbf{G}_{M+1}^{(l)}(k)$ is used. Then

$$\mathbf{G}_{M+1}^{(1)}(k) \times \begin{bmatrix} \lambda^{0.5}\hat{\mathbf{R}}_N(k-1) \\ \tilde{\mathbf{Y}}_M^{(0)H}(k) \end{bmatrix} = \begin{bmatrix} c_{M,1}(k) & 0 & \cdots & 0 & -S_{M,1}^*(k) \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \mathbf{I}_{N-3} & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ s_{M,1}(k) & 0 & \cdots & 0 & c_{M,1}(k) \end{bmatrix} \times$$

$$\times \begin{bmatrix} \lambda^{0.5}\hat{R}_{M,11}(k-1) & \lambda^{0.5}\hat{R}_{M,12}(k-1) & \cdots & \lambda^{0.5}\hat{R}_{M,1(M-1)}(k-1) & \lambda^{0.5}\hat{R}_{M,1M}(k-1) \\ 0 & \lambda^{0.5}\hat{R}_{M,22}(k-1) & \cdots & \lambda^{0.5}\hat{R}_{M,2(M-1)}(k-1) & \lambda^{0.5}\hat{R}_{M,2M}(k-1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \hat{R}_{M,(M-1)(M-1)}(k-1) & \hat{R}_{M,(M-1)M}(k-1) \\ 0 & 0 & \cdots & 0 & \hat{R}_{M,MM}(k-1) \\ \tilde{y}_{M,1}^{(0)*}(k) & \tilde{y}_{M,2}^{(0)*}(k) & \cdots & \tilde{y}_{M,(M-1)}^{(0)*}(k) & \tilde{y}_{M,M}^{(0)*}(k) \end{bmatrix} =$$

$$= \begin{bmatrix} \lambda^{0.5}\hat{R}_{M,11}(k) & \lambda^{0.5}\hat{R}_{M,12}(k-1) & \cdots & \lambda^{0.5}\hat{R}_{M,1(M-1)}(k-1) & \lambda^{0.5}\hat{R}_{M,1M}(k-1) \\ 0 & \lambda^{0.5}\hat{R}_{M,22}(k-1) & \cdots & \lambda^{0.5}\hat{R}_{M,2(M-1)}(k-1) & \lambda^{0.5}\hat{R}_{M,2M}(k-1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \hat{R}_{M,(M-1)(M-1)}(k-1) & \hat{R}_{M,(M-1)M}(k-1) \\ 0 & 0 & \cdots & 0 & \hat{R}_{M,MM}(k-1) \\ 0 & \tilde{y}_{M,2}^{(1)*}(k) & \cdots & \tilde{y}_{M,(M-1)}^{(1)*}(k) & \tilde{y}_{M,M}^{(1)*}(k) \end{bmatrix}. \tag{19}$$

In equation (19), vector $\tilde{\mathbf{Y}}_N^{(0)H}(k)$ is defined as

$$\tilde{\mathbf{Y}}_N^{(0)H}(k) = [\tilde{y}_{M,1}^{(0)*}(k), \tilde{y}_{M,2}^{(0)*}(k), \ldots, \tilde{y}_{M,M}^{(0)*}(k)] = \mathbf{Y}_N^H(k). \tag{20}$$

The superscript in parentheses means the number of the transformation performed on variable $\tilde{y}_{M,i}^{(0)*}(k)$ with number i in the vector. Such a conversion over $\tilde{y}_{M,1}^{(0)*}(k)$ is performed once (after the first time the variable is reset), over $\tilde{y}_{M,2}^{(0)*}(k)$ — twice (after the second time, the variable is reset), etc. and over $\tilde{y}_{M,M}^{(0)*}(k)$ — M times (after the M-th time, the variable is reset).

From equation (19) it follows that

$$c_{M,1}(k)\lambda^{0.5}\hat{R}_{M.11}(k-1) - s_{M,1}^*(k)\tilde{y}_{M,1}^{(0)*}(k)(k) = \hat{R}_{M.11}(k) \tag{21}$$

and

$$s_{M,1}(k)\lambda^{0.5}\hat{R}_{M.11}(k-1) + c_{M,1}(k)\tilde{y}_{M,1}^{(0)*}(k) = 0. \tag{22}$$

Performing similar conversion of all $i = 1, 2, \ldots, M$, can be established that

$$c_{M,i}(k)\lambda^{0.5}\hat{R}_{M.ii}(k-1) - s_{M,i}^*(k)\tilde{y}_{M,i}^{(i-1)*}(k) = \hat{R}_{M.ii}(k), \tag{23}$$

$$s_{M,i}(k)\lambda^{0.5}\hat{R}_{M.ii}(k-1) + c_{M,i}(k)\tilde{y}_{M,i}^{(i-1)*}(k) = 0. \tag{24}$$

Then from equation (22) we can determine that

$$s_{M,i}(k) = -c_{M,i}(k)\tilde{y}_{M,i}^{(i-1)*}(k)[\lambda^{0.5}\hat{R}_{M,ii}(k-1)]^{-1},$$

and, given that $c^2 + ss^* = 1$, from the equation

$$\begin{aligned} &c^2_{M,i}(k) + c_{M,i}(k)\tilde{y}^{(i-1)*}_{M,i}(k)[\lambda^{0.5}\hat{R}_{M.ii}(k-1)]^{-1}c_{M,i}(k)\tilde{y}^{(i-1)*}_{M,i}(k)[\lambda^{0.5}\hat{R}_{M.ii}(k-1)]^{-1} = \\ &= c^2_{M,i}(k)[\lambda^{0.5}\hat{R}_{M.ii}(k-1)]^2 + c^2_{M,i}(k)[\tilde{y}^{(i-1)*}_{M,i}(k)\tilde{y}^{(i-1)}_{M,i}(k)] = \\ &= c^2_{M,i}(k)[\lambda\hat{R}^2_{M,ii}(k-1) + \tilde{y}^{(i-1)*}_{M,i}\tilde{y}^{(i-1)}_{M,i}][\lambda\hat{R}^2_{M.ii}(k-1)]^{-1} = 1 \end{aligned} \tag{25}$$

can determine that

$$\begin{aligned} c_{M,i}(k) =& \sqrt{\lambda\hat{R}^2_{M.ii}(k-1)\left[\lambda\hat{R}^2_{M.ii}(k-1) + \tilde{y}^{(i-1)*}_{M,i}(k)\tilde{y}^{(i-1)}_{M,i}(k)\right]^{-1}} = \\ =& \sqrt{\lambda\hat{R}^2_{M.ii}(k-1)}\sqrt{\left[\lambda\hat{R}_{M.ii}(k-1) + \tilde{y}^{(i-1)*}_{M,i}(k)\tilde{y}^{(i-1)}_{M,i}(k)\right]^{-1}} = \lambda^{0.5}\hat{R}_{M.ii}(k-1)\hat{R}^{-1}_{M.ii}(k), \end{aligned} \tag{26}$$

where

$$\hat{R}_{M.ii}(k) = \sqrt{\lambda\hat{R}^2_{M.ii}(k-1) + \tilde{y}^{(i-1)*}_{M,i}(k)\tilde{y}^{(i-1)}_{M,i}(k)}. \tag{27}$$

It is taken into account that the diagonal elements $\hat{R}_{M.ii}(k)$ of the matrix $\hat{\mathbf{R}}_M(k)$ are real numbers. Then, using (26) in equation (24), we can determine

$$\begin{aligned} s_{N,i}(k) = -\lambda^{0,5}\tilde{R}_{M.ii}(k-1)\tilde{R}^{-1}_{M.ii}(k) + \tilde{y}^{(i-1)*}_{M,i}(k)\left[\lambda^{0.5}\hat{R}^2_{M.ii}(k-1)\right]^{-1} = \\ = -\tilde{y}^{(i-1)*}_{M,i}(k)\tilde{R}^{-1}_{M.ii}(k). \end{aligned} \tag{28}$$

Thus, equations (26)–(28) allow us to calculate the cosine and sine of a certain rotation angle. According to (19), they calculate the elements of the i-th row of matrix $\hat{\mathbf{R}}_M(k)$, zero out the i-th element of vector $\tilde{\mathbf{Y}}^{(i-1)H}_M(k)$, and modify the remaining nonzero elements of this vector as $\left[0, 0, \ldots, 0, 0, \tilde{y}^{(i)*}_{M,i+1}(k), \ldots, \tilde{y}^{(i)*}_{M,M}(k)\right]$, i.e.

$$\tilde{\mathbf{Y}}^{(i-1)H}_M(k) = \left[0, 0, \ldots, 0, \tilde{y}^{(i-1)*}_{M,i}(k), \tilde{y}^{(i-1)*}_{M,i+1}(k), \ldots, \tilde{y}^{(i-1)*}_{M,M}(k)\right] \tag{29}$$

and

$$\tilde{\mathbf{Y}}^{(i)H}_M(k) = \left[0, 0, \ldots, 0, 0, \tilde{y}^{(i)*}_{M,i+1}(k), \ldots, \tilde{y}^{(i)*}_{M,M}(k)\right]. \tag{30}$$

These transformations for each value of i include calculations (26)–(28), and for all $j = i+1, \ldots, M$, calculations similar to (23) and (24), i.e.

$$\hat{R}_{M.ij}(k) = c_{M,i}(k)\lambda^{0.5}\hat{R}_{M.ij}(k-1) - s^*_{M,i}(k)\tilde{y}^{(i-1)*}_{M,j}(k) \tag{31}$$

and

$$\tilde{y}^{(i)*}_{M,j}(k) = s_{M,i}(k)\lambda^{0.5}\hat{R}_{M.ij}(k-1) + c_{M,i}(k)\tilde{y}^{(i-1)*}_{M,j}(k). \tag{32}$$

Thus, transformation (13) can be performed either as $\mathbf{Q}^H_k(k)\mathbf{\Lambda}^{0.5}_k(k)\mathbf{Y}^H_{Mk)}(k)$, applying a $k \times k$ matrix $\mathbf{Q}^H_k(k)$ to a $k \times M$ matrix $\mathbf{\Lambda}^{0.5}_k(k)\mathbf{Y}^H_{Mk)}(k)$ at each iteration, or using a matrix $\hat{\mathbf{Q}}^H_k(k) = \prod_{i=1}^{M}\mathbf{G}^{(i)}_M(k)$ applied to a $k \times M$ matrix $\begin{bmatrix}\lambda^{0.5}\hat{\mathbf{R}}_M(k-1)\\ \mathbf{0}_{(k-1-M)M}\\ \mathbf{Y}^H_M(k)\end{bmatrix}$.

2. Parallel implementation of the QR algorithm in a triangular systolic array

Givens transformation has good properties for use in a triangular systolic array. The architecture of the basic computations of the algorithm using such calculators is given in [16] and in Fig. 1.

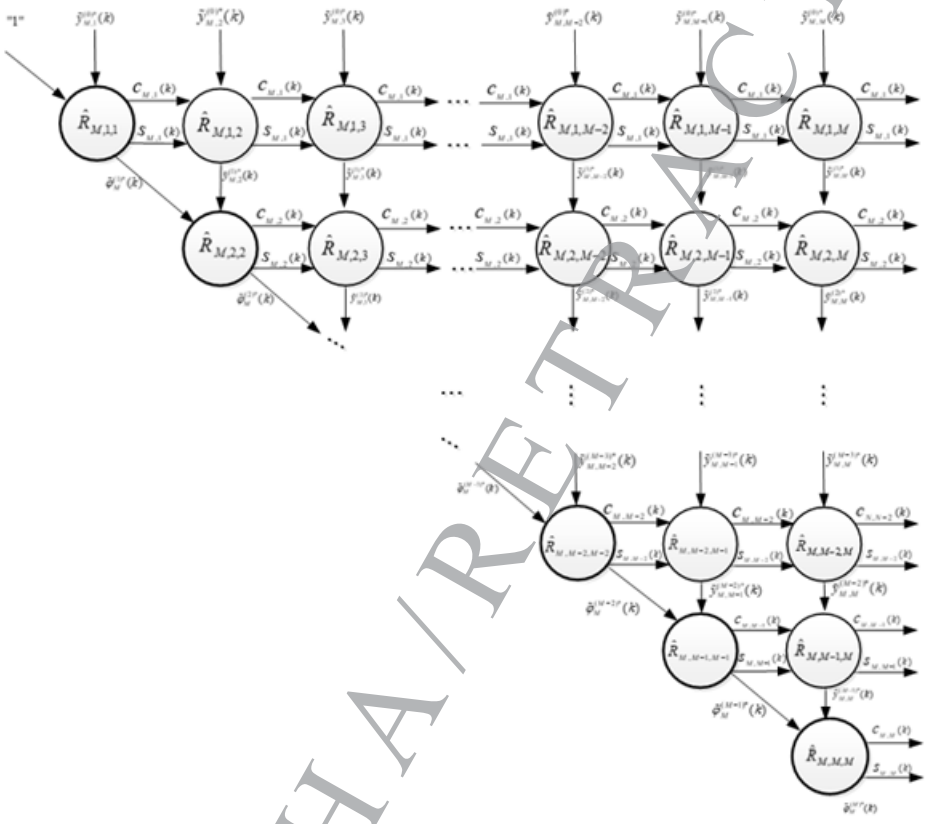

Fig. 1. Block diagram of a triangular systolic array

The systolic array is based on the method of triangular complex rotations and allows to obtain a significant performance gain in comparison with the method of complex rotations of Givens.

In the system of a triangular systolic array, there are individual processing cells located in an ordered structure. Each individual cell of the system has its own processing functionality and local memory. Only neighboring cells are connected to each other and there is no direct connection between cells that are not adjacent. When data is fed into the systolic array system, the processing cells at the front end of the system will process the data, store them in local memory, and then forward them to adjacent cells. This processing and transfer of the processed data in each cell continues until the data stream reaches the end of the system, where the final calculation results are obtained. The proposed architecture provides a significant reduction in the time required to perform QR decomposition using the same computing resources (CORDIC computational cells). Another advantage of the proposed scheme is that during QR decomposition, the upper triangular matrix $\mathbf{R}$ has only real diagonal elements. This simplifies the subsequent inversion of the matrix $\mathbf{R}$ using the backward substitution algorithm, which requires division by the diagonal elements of the matrix $\mathbf{R}$.

Algorithm (15) shows that the application of Givens rotations in a post-array to multiply matrices on the left side of the preliminary array allows one to obtain a triangular matrix and zero out the input vector $\mathbf{Y}(k)$. The number of elements in the input vector $\mathbf{Y}(k)$ corresponds to the number of antennas. The number of required Givens rotation operations is the same as the number of elements in the input. That is, each rotation of Givens will reset exactly one element of the input vector. Thus, the algorithm for generating a radiation pattern with K antennas requires the Givens rotation operation K in the calculation of the post-array.

In calculations, the Givens rotation operation can be performed in parallel, because there is no data dependence between the Givens rotation operation at one of the inputs and the Givens rotation operation at the same position in subsequent iterations. Thus, Givens rotation operations can be performed in parallel.

Conclusion

The proposed architecture of the triangular systolic array using the method of the triangular complex rotations optimized for implementation in large-scale integrated circuits, allowing you to effectively perform the operation QR-decomposition of complex matrices. Compared with the QR-RLS algorithm, the proposed architecture can provide a gain of up to 35% in the time of calculating the QR decomposition. The synthesized algorithm will make it possible to implement the methods of spatio-temporal processing of broadband signals of satellite communication systems.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation in the framework of the Federal target program «Research and development on priority directions of development of the scientific-technological complex of Russia for 2014-2020» (agreement no. 05.605.21.0185, unique ID project RFMEFI60519X0185).

References

[1] Ya.D.Shirman, Radio-electronic systems: Fundamentals of construction and theory. Directory. Ed. 2nd, rev. and add., Moscow, Radio Engineering, 2007 (in Russian).

[2] V.N.Tyapkin, I.N.Kartsan, D.D.Dmitriev, S.V.Efremova, Algorithms for adaptive processing of signals in a flat phased antenna array, SIBCON, 2017.

[3] A.I.Perov, Statistical theory of radio systems, Moscow, Radio Engineering, 2003 (in Russian).

[4] I.N.Kartsan, V.N.Tyapkin, D.D.Dmitriev, A.E.Goncharov, P.V.Zelenkov, I.V.Kovalev, *IOP Conf. Series: Materials Science and Engineering*, **155**(2016) no. 1. DOI: 10.1088/1757-899X/155/1/012019

[5] I.N.Kartsan, V.N.Tyapkin, D.D.Dmitriev, A.E.Goncharov, I.V.Kovalev, *IOP Conf. Series: Materials Science and Engineering*, **255**(2017). DOI: 10.1088/1757-899X/255/1/012009

[6] V.N.Tyapkin, D.D.Dmitriev, Yu.L.Fateev, N.S.Kremez, *J. Sib. Fed. Univ. Math. & Phys.*, **9**(2016), no. 2, 258–268. DOI: 10.17516/1997-1397-2016-9-2-258-267

[7] I.N.Kartsan, Y.L.Fateev, V.N.Tyapkin, D.D.Dmitriev, A.E.Goncharov, P.V.Zelenkov, I.V.Kovalev, *IOP Conf. Series: Materials Science and Engineering,* **155**(2016), no. 1. DOI: 10.1088/1757-899X/155/1/012020

[8] R.A.Monzingo, T.W.Miller, Introduction to Adaptive Arrays, New York, John Wiley & Sons, 1980.

[9] Ya.D.Shirman et al., The first domestic studies of the adaptation of antenna systems to interfering influences, Moscow, Radio engineering, no. 11, 1989 (in Russian).

[10] V.N.Tyapkin, D.D.Dmitriev, V.G.Konnov, A.N.Fomin, A method for determining the vector of spectral coefficients by the likelihood ratio criterion, *Bulletin of the Siberian state Aerospace University named after Acad. M. F. Reshetneva*, **43**(2012), no. 3, 76–79.

[11] V.I.Dzhigan, Equivalence conditions for recursive adaptive filtering algorithms by the least squares criterion, *Telecommunications*, **6**(2006), 6–11 (in Russian).

[12] I.A.Lubkin, V.N.Tyapkin, The use of recurrent adaptive algorithms to solve the problem of suppressing active noise interference in satellite communication systems, *Bull. Sib. state Aerospace Univ. named after Acad. M.F. Reshetneva*, **28**(2010), no. 2, 39–43 (in Russian).

[13] G.H.Golub, C.F.Van, Lone Matrix calculations, The Johns Hopkins University Press, Baltimor, 1996.

[14] D.T.M.Slock, T.Kailath, Numerically stable fast transversal filters for recursive least squares adaptive filtering, *IEEE Trans. Signal Processing*, **39**(1991), no. 1, 92–114.

[15] A.Benallal, A.Gilliore, A new method to stabilize fast RLS algorithm based on the first-order model of the propagation of numerical errors, Proc. Int. Conf. on Acoustic, Speech and Signal Processing, Vol. 5, 1988, 1373–1376.

[16] J.G.McWhirter, Recursive least-squares minimization using a systolic array, Proceedings of the SPIE Intern. Sic. Opt. Eng. Vol. 43, 1983, 105–112.

Рекурсивный алгоритм оценивания корреляционной матрицы помех, основанный на QR-разложении

Валерий Н. Тяпкин
Дмитрий Д. Дмитриев
Андрей Б. Гладышев
Пётр Ю. Зверев
Сибирский федеральный университет
Красноярск, Российская Федерация

Аннотация. Многие задачи цифровой обработки сигналов требуют выполнения матричных операций в режиме реального времени. Это операции обращения матрицы или решения систем линейных алгебраических или дифференциальных уравнений (фильтр Калмана). Переход к реализации цифровой обработки сигналов на программируемых логических интегральных схемах (ПЛИС), как правило, предполагает расчеты, основанные на представлении чисел с фиксированной точкой. Это делает практически невозможным решение задач пространственно-временной обработки на основе традиционных вычислительных методов. В статье рассматривается реализация алгоритмов пространственно-временной обработки сигналов в широкополосных спутниковых системах с использованием QR-разложения. Представлены технологии вычислений CORDIC, необходимые для повторного QR-разложения при совместном использовании в систолических алгоритмах.

Ключевые слова: фазированная антенная решетка, адаптивные алгоритмы, фильтр Калмана, рекурсивный алгоритм по критерию наименьших квадратов, QR-разложение, систолические алгоритмы.

DOI: 10.17516/1997-1397-2020-13-2-170-186
УДК 512.55+517.95

Ideals Generated by Differential Equations

OlegV. Kaptsov*
Institute of computational modelling SB RAS
Krasnoyarsk, Russian Federation

Received 13.11.2019, received in revised form 22.01.2020, accepted 06.02.2020

Abstract. We propose a new algebraic approach to study compatibility of partial differential equations. The approach uses concepts from commutative algebra, algebraic geometry and Gröbner bases to clarify crucial notions concerning compatibility such as passivity and reducibility. One obtains sufficient conditions for a differential system to be passive and proves that such systems generate manifolds in the jet space. Some examples of constructions of passive systems associated with the sinh-Cordon equation are given.

Keywords: differential rings and ideals, Gröbner bases, partial differential equations.

Citation: O.V.Kaptsov, Ideals Generated by Differential Equations, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 170–186. DOI: 10.17516/1997-1397-2020-13-2-170-186.

Introduction

There are currently no methods to study the general systems of partial differential equations. Therefore it is necessary to investigate special classes of equations. For example, the linear systems of homogeneous first order differential equations with one unknown function form one of well-studied classes [1, 2].

In the beginning of the twentieth century, French mathematicians Riquier, Janet, and Cartan made significant progress in studying a broad class of partial differential equations [3–5]. Over the past several decades, new tools and terminology coming from differential geometry, differential and commutative algebra began to be applied in the formal theory of differential equations [6–8]. It is now becoming increasingly important to consider algorithmic problems of the theory of differential equations [9, 10]. Some algorithms are implemented in computer algebra systems such as Maple, Reduce, Mathematica.

In the papers [11, 12], we used tools from the algebraic geometry and Gröbner bases to study local properties of analytic partial differential equations. Here we consider smooth case. Some our notions can be explained by means of an example. Consider the $n+m$-dimensional real space $\mathbb{R}^{n+m}$ equipped with the natural coordinate functions $x_1, \dots, x_n, y_1, \dots, y_m$ and the standard topology. Denote by $\mathfrak{F}(V)$ the algebra of smooth functions on an open set $V \subset \mathbb{R}^{n+m}$ and denote by $\mathfrak{F}_a$ the algebra of germs of smooth functions at a point $a \in \mathbb{R}^{n+m}$. A subset $B = \{f_1, \dots, f_m\}$ of $\mathfrak{F}(V)$ is called a normalized set, if each function $f_i \in B$ is of the form

$$f_i = y_i + g_i(x),$$

where the function g_i can depends only on $x_1, \dots, x_n$. We say that an ideal of the algebra $\mathfrak{F}(V)$ is soft if it is generated by an normalized set. It is easy to give analogous definitions in the case of the algebra $\mathfrak{F}_a$.

The goal of this paper is to present an algebraic technique for studying compatibility of smooth partial differential equations. Section 1 deals with the infinite-dimensional space $\mathbb{R}^T$ of

*kaptsov@icm.krasn.ru

all the maps $T \longrightarrow \mathbb{R}$ equipped with the product topology (where T is a countable set). To each open set V of the space $\mathbb{R}^T$ one associates an algebra $\mathcal{F}(V)$ of smooth functions on V depending only on finitely many variables. The set of all germs of these functions at a point $a \in \mathbb{R}^T$ forms a local algebra $\mathcal{F}_a$. Next we define the appropriate normalized sets and soft ideals in the algebras $\mathcal{F}(V)$ and $\mathcal{F}_a$. It turns out that every normalized set leads to a manifold in $\mathbb{R}^T$.

Let $\mathbb{N}$ be the set of all non-negative integers and $\mathbb{N}_k = \{1, \dots, k\}$. In Section 2 we consider the infinite jet space $\mathbb{J} = \mathbb{R}^T$ with $T = \mathbb{N}_n \cup (\mathbb{N}_m \times \mathbb{N}^n)$, then a system of partial differential equations is a subset of the algebra $\mathcal{F}(V)$. We define passive systems of partial differential equations at a point and on an open set in $\mathbb{J}$. These notations are analogous to Gröbner bases [13], but our definition does not apply any ranking.

In Section 3 we introduce the basic tools for study passive systems. One of these is a stratified set which is given by a partition and a monoid acting on the set. Any stratified set must satisfy certain compatibility conditions. The monoid $(\mathbb{N}^n, +)$ acts on the algebras $\mathcal{F}(V)$ and $\mathcal{F}_a$ by means of derivations. The stratification allows us to introduce reductions of functions as well as reductions of germs modulo differential systems and to define reducibility conditions at a point and on an open set in $\mathbb{J}$.

The crucial theorems are given in Section 4. We prove that if a differential system S is a normalize set and satisfies reducibility conditions at a point, then it generates a soft ideal and it is passive. Furthermore, if the system satisfies reducibility conditions on an open set, then the orbit of S leads to a manifold in the infinite jet space $\mathbb{J}$. At the end of our paper we give examples of passive systems dealing with sinh-Gordon equation.

1. Normalized sets in an algebra of smooth functions

We shall use the following notations $\mathbb{R}$, for the set of all real numbers, $\mathbb{N}$, for the set of all non-negative integer, $\mathbb{N}_k$, for the set $\{1, 2, \dots, k\}$. Let T be a denumerable set; the space of maps $z : T \longrightarrow \mathbb{R}$ is denoted as $\mathbb{R}^T$ and equipped with the product topology. In this case a neighbourhood base for any point $a \in \mathbb{R}^T$ is given by

$$U(a_\tau, \rho) = \{z \in \mathbb{R}^T : |z_{t_i} - a_{t_i}| < \rho_i,\ i \in \mathbb{N}_k\}, \tag{1.1}$$

where $t_i \in T,\ \rho_i > 0,\ \rho = (\rho_1, \dots, \rho_k),\ a_\tau = \{a_{t_1}, \dots, a_{t_k}\}$ is a set of k coordinates of the point a; $z_{t_1}, \dots, z_{t_k}$ are k coordinates of the point z. The functions $y_t : \mathbb{R}^T \longrightarrow \mathbb{R}$ defined by $y_t(z) = z(t),\ t \in T$, are *the standard coordinate functions* (variables). The set $Y = \{y_t\}_{t \in T}$ is *the standard coordinate system* on $\mathbb{R}^T$.

Let V be an open set in $\mathbb{R}^T$ and let $\mathcal{F}(V)$ be the $\mathbb{R}$-algebra of real functions on V that depend on finitely many variables and are smooth (i.e. they have derivatives of all orders) as functions of a finite number of variables. Suppose a function $f \in \mathcal{F}(V)$ depends on some set of variables, then this set denotes by ivf. When H is a subset of $\mathcal{F}(V)$, we shall use the notation

$$ivH = \{ivf : f \in H\}. \tag{1.2}$$

The family $\{\mathcal{F}(V)\}_{V \subset \mathbb{R}^T}$ gives rise to the sheaf $\mathcal{F}$ of smooth functions on $\mathbb{R}^T$. For each point $a \in \mathbb{R}^T$ a stalk $\mathcal{F}_a$ of the sheaf is a $\mathbb{R}$-algebra of germs of smooth functions at a. Given a function $f \in \mathcal{F}(V)$, then its germ at a is denoted as $\tilde{f}_a$ or $\tilde{f}$ for simplicity.

Each stalk $\mathcal{F}_a$ of the sheaf $\mathcal{F}$ is a local algebra. Indeed, if $\tilde{f} \in \mathcal{F}_a$ and $\tilde{f}(a) \neq 0$, then $1/\tilde{f} \in \mathcal{F}_a$ and $\tilde{f}$ does not belong to any proper ideal of the algebra. Hence the set

$$\mathfrak{M}_a = \{\tilde{f} \in \mathcal{F}_a : \tilde{f}(a) = 0\} \tag{1.3}$$

is a unique maximal ideal of $\mathcal{F}_a$.

We shall say that a germ $\widetilde{f} \in \mathcal{F}_a$ depends on $\tilde{y}_t$ if there is a neighborhood V of a such that any representative f of $\tilde{f}$ depends on y_t in every neighborhood $V' \subset V$ of a. Assume a germ $\tilde{f} \in \mathcal{F}_a$ depends on a set of variables, then this set denotes by $iv\tilde{f}$.

Definition 1.1. *2.1. A set $B \subset \mathcal{F}(V)$ is called normalized if the following conditions hold:*

(i) any function $f \in B$ can be written $f = y_t + g$, where the coordinate functions y_t form a set $\mathcal{L}$ and the functions g do not depend on elements of $\mathcal{L}$;

(ii) if $f_1 = y_t + g_1$, $f_2 = y_t + g_2 \in B$, then $f_1 = f_2$. The elements of the sets $\mathcal{L}$ and $Y \setminus \mathcal{L}$ are called principal and parametric respectively.

We shall give a similar definition for germs. Let $\tilde{Y_a}$ denote the set of germs of the coordinate functions at a.

Definition 1.2. *A set $\tilde{B} \subset \mathcal{F}_a$ is called normalized if the following conditions hold:*

(i) every germ $\tilde{f} \in \tilde{B}$ can be written $\tilde{f} = \tilde{y}_t + \tilde{g}$, where the germs $\tilde{y}_t$ form a set $\tilde{\mathcal{L}} \subset \tilde{Y_a}$ and the germs $\tilde{g}$ do not depend on elements of $\tilde{\mathcal{L}}$;

(ii) if $\tilde{f}_1 = \tilde{y}_t + \tilde{g}_1$, $\tilde{f}_2 = \tilde{y}_t + \tilde{g}_2 \in \tilde{B}$, then $\tilde{f}_1 = \tilde{f}_2$. The elements of the set $\tilde{\mathcal{L}}$ are called the principal variables and elements of the set $\tilde{Y_a} \setminus \tilde{\mathcal{L}}$ are parametric variables.

Proposition 1.3. *Suppose $\tilde{f}_i = \tilde{y_{t_i}} + \tilde{g}_i$, $i \in \mathbb{N}_k$, are some elements of a normalized set $\tilde{B} \subset \mathcal{F}_a$ and a germ $\tilde{F} \in \mathcal{F}_a$ depends on $\tilde{y_{t_1}}, \ldots, \tilde{y_{t_k}}$. Then there exist germs $\tilde{q_1}, \ldots, \tilde{q_k} \in \mathcal{F}_a$ and a unique germ $\tilde{r} \in \mathcal{F}_a$ which does not depend on $\tilde{y_{t_1}}, \ldots, \tilde{y_{t_k}}$ such that*

$$\tilde{F} = \sum_{i=1}^{k} \tilde{q}_i \tilde{f}_i + \tilde{r}. \tag{1.4}$$

Proof. Suppose the germs $\tilde{F}, \tilde{f}_1, \ldots, \tilde{f}_k$ depend on $\tilde{y_{t_1}}, \ldots, \tilde{y_{t_n}}$. From the Mather division theorem [15], we obtain

$$\tilde{F} = \tilde{q_1} \tilde{f}_1 + \widetilde{r}_1,$$

where $\widetilde{q}_1 \in \mathcal{F}_a$; $\widetilde{r}_1 \in \mathcal{F}_a$ does not depend on $\widetilde{y}_{t_1}$. Using this theorem to the germ $\widetilde{r}_1$ yields

$$\tilde{F} = \tilde{q_1} \tilde{f}_1 + \tilde{q_2} \tilde{f}_2 + \widetilde{r}_2,$$

where $\widetilde{r}_2$ does not depend on $\widetilde{y}_{t_1}, \widetilde{y}_{t_2}$. Continuing in the same way, we derive (1.4).

One needs to verify uniqueness of $\tilde{r}$. Assume there exists another representation of $\tilde{F}$

$$\tilde{F} = \sum_{i=1}^{k} \tilde{q_i}^{\star} \tilde{f}_i + \tilde{r}^{\star}. \tag{1.5}$$

It follows from (1.4) and (1.5) that

$$\tilde{f} = \sum_{i=1}^{k} \tilde{h_i} \tilde{f}_i,$$

with $\tilde{f} = \tilde{r}^{\star} - \tilde{r}$, $\tilde{h_i} = \tilde{q_i} - \tilde{q_i}^{\star}$. Let f, h_i, f_i be representatives of the germs $\tilde{f}, \tilde{h_i}, \tilde{f}_i$. Then there is a neighborhood of a such that

$$f = \sum_{i=1}^{k} h_i f_i. \tag{1.6}$$

Next we introduce new variables

$$y'_{t_1} = f_1, \ldots, y'_{t_k} = f_k. \tag{1.7}$$

Since $f_i = y_{t_i} + g_i$ in some neighborhood of a, we can find y_{t_i} from (1.7) and substitute in the expression (1.6). Then we may write

$$f = \sum_{i=1}^{k} \bar{h_i} y'_{t_i},$$

where $\bar{h_1}, \ldots, \bar{h_k}$ are some smooth functions while f can only depend on $y_{t_{k+1}}, \ldots, y_{t_n}$. Assuming that

$$y'_{t_1} = 0, \ldots, y'_{t_k} = 0,$$

we have $f = 0$ and therefore $\tilde{r} = \tilde{r}^\star$. □

Proposition 1.4. *Let $B \subset \mathcal{F}(V)$ be a normalized set. Assume that a function $F \in \mathcal{F}(V)$ is a polynomial in some principal variables $y_{t_1}, \ldots, y_{t_k}$ of B with coefficients depending only on parametric variables. Then there is a unique function $r \in \mathcal{F}(V)$ not depending on the principal variables and some functions $q_1, \ldots, q_k \in \mathcal{F}(V)$ such that*

$$F = \sum q_i f_i + r, \tag{1.8}$$

where $f_i = y_{t_i} + g_i \in B$.

Proof. The function F is a polynomial in the principal variables $y_{t_1}, \ldots, y_{t_k}$ and the functions $f_1, \ldots, f_k$ are polynomials of the first degree with coefficients 1. Then we can obtain (1.8) using the multivariate division with remainder [13], although $\mathcal{F}(V)$ is not a field. Moreover, the function r does not depend on the principal variables and lies in $\mathcal{F}(V)$.

The uniqueness of r can be proved as in Proposition 1.3. Suppose that the function F is written in the other form

$$F = \sum q'_i f_i + r', \tag{1.9}$$

where the function r' does not depend on the principal variables. Then from (1.8) and (1.9) we have

$$r'' = \sum q''_i f_i \tag{1.10}$$

with $r'' = r - r'$ and $q''_i = q'_i - q_i$. Under the transformation

$$y'_{t_1} = f_1, \ldots, y'_{t_k} = f_k. \tag{1.11}$$

the relation (1.10) becomes

$$r'' = \sum q^\star_i y'_{t_i},$$

where $q^\star_1, \ldots, q^\star_k \in \mathcal{F}(V)$, while the function r'' does not depend on $y_{t_1}, \ldots, y_{t_k}$. Setting

$$y'_{t_1} = 0, \ldots, y'_{t_k} = 0,$$

we obtain $r'' = 0$. □

Remark. Inserting the values $y_{t_1} = -g_1, \ldots, y_{t_k} = -g_k$ in the function F, we obtain the function r.

A general definition of a smooth manifold is in [14], but we shall only consider embedded submanifolds of $\mathbb{R}^T$.

Definition 1.5. *Let V be an open set in $\mathbb{R}^T$. A map $\phi : V \to \mathbb{R}^T$ is called smooth on V if for all $t \in T$ the functions $\phi_t = y_t \circ \phi$ are smooth on V.*

Let V, V' be open sets in $\mathbb{R}^T$. We say that a map $\psi : V \to V'$ is a diffeomorphism if ψ carries V homeomorphically onto V' and if ψ and ψ^{-1} are smooth. If $T^\star \subset T$, then a set

$$C_{T^\star} = \{z \in \mathbb{R}^T : z(t) = 0, \forall t \in T\}$$

is called a coordinate subspace of $\mathbb{R}^T$. We shall assume that any subset $Q \subset \mathbb{R}^T$ is equipped with a topology induced from that of $\mathbb{R}^T$.

Definition 1.6. *A subset $M \subset \mathbb{R}^T$ is called a smooth manifold if for any $a \in M$ there are a neighborhood $V \subset \mathbb{R}^T$, an open subset $V' \subset \mathbb{R}^T$, and a diffeomorphism $\phi : V \to V'$ such that*

$$\phi(V \cap M) = V' \cap C_{T^\star},$$

where $C_{T^\star}$ is a coordinate subspace of $\mathbb{R}^T$.

Proposition 1.7. *Assume that $\{g_t\}_{t\in T'}$ is a family of smooth function on an open subset $W \subset \mathbb{R}^{T''}$ with $T'' = T \setminus T'$ and denote by V the open set $W \times \mathbb{R}^{T'}$ in $\mathbb{R}^T$. Then the set $B = \{y_t + g_t\}_{t\in T'} \subset \mathcal{F}(V)$ is normalized and the set*

$$Z(B) = \{z \in V : f(z) = 0, f \in B\}$$

is a manifold in $\mathbb{R}^T$.

Proof. Let $\phi : V \to \mathbb{R}^T$ be a map given by

$$y'_t = y_t + g_t, \qquad y'_s = y_s \qquad \forall t \in T' \forall s \in T''.$$

Then the inverse map is of the form

$$y_t = y'_t - g_t, \quad y_s = y'_s.$$

It is easy to see that

$$\phi(V \cap Z(B)) = V \cap \mathbb{R}^{T''},$$

and hence $Z(B)$ is a manifold. □

2. Passive differential systems

We now introduce the basic notions concerning compatibility of partial differential equations.

Definition 2.1. *(i) We say that a proper ideal I of an algebra $\mathcal{F}(V)$ is soft if there is a normalized set $B \subset \mathcal{F}(V)$ to generate the ideal. The set B is called a normalized system of generators of I.*

(ii) Let J be a proper ideal of an algebra $\mathcal{F}_a$. A normalized subset $\tilde{B} \subset \mathcal{F}_a$ generating the ideal J is called a normalized system of generators of J and we say that the ideal is soft.

We recall that a derivation in an algebra A over $\mathbb{R}$ is a map $\mathcal{D} : A \longrightarrow A$ such that

$$\mathcal{D}(ab) = a\mathcal{D}(b) + \mathcal{D}(a)b, \quad \mathcal{D}(k_1a + k_2b) = k_1\mathcal{D}(a) + k_2\mathcal{D}(b)$$

for all $a, b \in A$ and for all $k_1, k_2 \in \mathbb{R}$.

The next proposition describes an arbitrary derivation of the algebra of germs $\mathcal{F}_a$.

Proposition 2.2. *Let $\mathcal{D}, \bar{\mathcal{D}}$ be derivations of the algebra $\mathcal{F}_a$ such that $\mathcal{D}(y_t) = \bar{\mathcal{D}}(y_t)$ for all $y_t \in Y$. Then $\mathcal{D} = \bar{\mathcal{D}}$ and*

$$\mathcal{D}(\tilde{f}) = \sum_{t\in T} \frac{\partial \tilde{f}}{\partial \tilde{y}_t} \mathcal{D}(\tilde{y}_t), \quad \forall \tilde{f} \in \mathcal{F}_a. \tag{2.1}$$

Proof. Repeating the proof Theorem 4.2 (a variant of Hadamard's lemma) in [15], we see that the set $\tilde{Y}$ of germs $\{\tilde{y_t}\}_{t\in T}$ at $a \in \mathbb{R}^T$ generates the maximal ideal (1.3). It follows from the Proposition 8.16 [16] that $\mathcal{D} = \bar{\mathcal{D}}$. It is easy to see that the expression (2.1) gives the derivations of $\mathcal{F}_a$. Even though the formula (2.1) involves an infinity summation, when applying $\mathcal{D}$ to any germ $\tilde{f}$, only finitely many terms are need.

Now we proceed to consider differential equations. Further, assume that

$$T = \mathbb{N}_n \cup (\mathbb{M} \times \mathbb{N}^n),$$

where $\mathbb{M} = \mathbb{N}_m$ or $\mathbb{M} = \mathbb{N}$. By $\mathbb{J}$ denote the space $\mathbb{R}^T$ and call it the *jet space*. The standard coordinate functions on $\mathbb{J}$ are denoted by $x_1, \dots, x_n$, u^i_α, where $i \in \mathbb{M}, \alpha \in \mathbb{N}^n$. The standard coordinate system Y on $\mathbb{J}$ is decomposed into two sets

$$X = \{x_1, \dots, x_n\}, \qquad U = \{u^i_\alpha\}^{i\in\mathbb{M}}_{\alpha\in\mathbb{N}^n}. \tag{2.2}$$

The elements $e_1 = (1, 0, \dots, 0), \dots, e_n = (0, \dots, 1)$ are generators of the monoid $\mathbb{N}^n$. Introduce derivations $\mathcal{D}_1, \dots, \mathcal{D}_n$ on the algebras $\mathcal{F}(V)$, $\mathcal{F}_a$ so that

$$\mathcal{D}_i f = \frac{\partial f}{\partial x_i} + \sum_{j\in\mathbb{M},\alpha\in\mathbb{N}^n} \frac{\partial f}{\partial u^i_\alpha} u^i_{\alpha+e_j}, \quad \mathcal{D}_i \tilde{f} = \frac{\partial \tilde{f}}{\partial \tilde{x_i}} + \sum_{j\in\mathbb{M},\alpha\in\mathbb{N}^n} \frac{\partial \tilde{f}}{\partial \tilde{u^i_\alpha}} \tilde{u^i}_{\alpha+e_j}. \tag{2.3}$$

Thus $\mathcal{F}(V)$ and $\mathcal{F}_a$ became differential algebras. □

We, following Ritt's terminology [17], call the coordinate functions u^i_0 the *indeterminates* and u^i_α the partial derivatives of u^i_0.

Definition 2.3. *We shall say that a subset $S \subset \mathcal{F}(V)$ is a differential system on an open set $V \subset \mathbb{J}$ if any function $f \in S$ depends on at least one of the partial derivatives. If $\mathbb{M} = \mathbb{N}_m$ then we say that S is a system with finite number of indeterminates, but if $\mathbb{M} = \mathbb{N}$ then we get a system in infinitely many indeterminates.*

Let W be an open set in $\mathbb{R}^n$ and let $h : W \to \mathbb{R}^{\mathbb{M}}$ be a smooth map with components h_m for $m \in \mathbb{M}$. Then a map $h^\infty : W \to \mathbb{J}$ whose components are $x_i, h^\alpha_m = \mathcal{D}^\alpha(h_m)$ for $i \in \mathbb{N}_n$, $m \in \mathbb{M}$, $\alpha \in \mathbb{N}^n$ is called the infinite prolongation graph of h.

Definition 2.4. *Let S be a differential system on an open subset $V \subset \mathbb{J}$. A smooth map $h : W \to \mathbb{R}^{\mathbb{M}}$ is called a solution of a differential system S if the following conditions hold:*

$$(1) \quad h^\infty(W) \subset V, \qquad (2) \quad f \circ h^\infty = 0, \quad \forall f \in S.$$

Remarks. In other words, the map h is a solution of the system S if under substitution of $\mathcal{D}^\alpha(h_i)$ for u^i_α every function $f \in S$ vanishes. A germ of a solution is defined in the obvious way.

An ideal of the algebra $\mathcal{F}(V)$ generated by a set $\{\mathcal{D}^\alpha(f) : f \in S, \alpha \in \mathbb{N}^n\}$ we shall denote by $\langle\langle S\rangle\rangle$. Similarly, let $\tilde{S}_a$ be a set of germs of functions in $S \subset \mathcal{F}(V)$ at a. An ideal of the algebra $\mathcal{F}_a$ generated by the set $\{\mathcal{D}^\alpha(\tilde{f}) : \tilde{f} \in \tilde{S}, \alpha \in \mathbb{N}^n\}$, denoted by $\langle\langle \tilde{S}\rangle\rangle_a$.

It is obvious that a map h is a solution of a differential system S if and only if $f \circ h^\infty = 0$ for all $f \in \langle\langle S\rangle\rangle$. There are some cases in which it is convenient to deal with other differential system S' such that $\langle\langle S'\rangle\rangle = \langle\langle S\rangle\rangle$. In particular, such examples arise when we consider compatible systems of differential equations of the first order for a single unknown function [2].

Recall that if G and H are sets, then G acts on H in case there is a mapping $\psi : G \times H \longrightarrow H$. The mapping ψ is called a action. When ψ is fixed, then gh denotes $\psi(g, h)$. The monoid $(\mathbb{N}^n, +, 0)$ acts on the algebras $\mathcal{F}(V)$, $\mathcal{F}_a$ by

$$\alpha f = \mathcal{D}^\alpha f, \qquad \alpha \tilde{f} = \mathcal{D}^\alpha \tilde{f}, \qquad \forall \alpha \in \mathbb{N}^n \forall f \in \mathcal{F}(V) \forall \tilde{f} \in \mathcal{F}_a.$$

The sets

$$O(f) = \{\mathcal{D}^\alpha f : \alpha \in \mathbb{N}^n\}, \quad O(\tilde{f}) = \{\mathcal{D}^\alpha \tilde{f} : \alpha \in \mathbb{N}^n\}$$

are orbits of a function f and a germ $\tilde{f}$ under $(\mathbb{N}^n, +, 0)$.

Definition 2.5. *(i) A germ $\tilde{f} \in \mathcal{F}_a$ of the form $\tilde{f} = \tilde{u}^i_\alpha + \tilde{g}$ is called solvable with respect to $\tilde{u}^i_\alpha$ if the germ $\tilde{g}$ does not depend on elements of the orbit $O(\tilde{u}^i_\alpha)$.*

(ii) A function $f = u^i_\alpha \in \mathcal{F}(V)$ is solvable with respect to u^i_α if the function g does not depend on elements of the orbit $O(u^i_\alpha)$.

Suppose a germ $\tilde{f} \in \mathcal{F}_a$ is solvable with respect to $\tilde{u}^i_\alpha$. Then the germ $\tilde{u}^i_\alpha$ is denoted by $st\tilde{f}$. Let $\tilde{S}_a$ be a set of solvable germs at a point a, then we shall use the notation $st\tilde{S} = \{st\tilde{f} : \tilde{f} \in \tilde{S}_a\}$. The same notation is used for functions.

Definition 2.6. *A differential system $S \subset \mathcal{F}(V)$ is called passive at $a \in V$ if the ideal $\langle\langle \tilde{S} \rangle\rangle_a$ is smooth, the set $\tilde{S}_a$ consists of solvable germs, and a set of principal variables of a normalized system of the ideal $\langle\langle \tilde{S} \rangle\rangle_a$ coincides with the orbit $O(st\tilde{S}_a)$. The system S is passive on V if every function in S is solvable, the ideal $I = \langle\langle S \rangle\rangle$ is smooth, and a set of principal variables of a normalized system of the ideal I coincides with the orbit $O(stS)$.*

3. Stratified sets and reductions

We need a convenient criterion for recognizing passive systems. For this purpose, we shall introduce additional tools. Recall that a preorder $\preceq$ is a binary relation that is reflexive and transitive. A strict partial order $\prec$ is a binary relation that is irreflexive and transitive.

In what follows, we shall deal with a well-ordered set Γ. Every partition $\{H_\gamma\}_{\gamma\in\Gamma}$ of a set H gives rise to a preorder and a strict partial order on H as follows:

$$h_1 \preceq h_2 \quad \Longleftrightarrow \quad \exists \gamma_1, \gamma_2 \in \Gamma (\gamma_1 \leqslant \gamma_2 \wedge h_1 \in H_{\gamma_1} \wedge h_2 \in H_{\gamma_2}), \tag{3.1}$$

$$h_1 \prec h_2 \quad \Longleftrightarrow \quad \exists \gamma_1, \gamma_2 \in \Gamma (\gamma_1 < \gamma_2 \wedge h_1 \in H_{\gamma_1} \wedge h_2 \in H_{\gamma_2}). \tag{3.2}$$

In this case we say that the set H is equipped with a induced strict partial order. We also say that a monoid G acts on the set H if there exists a map $(g, x) \longrightarrow gx$ of $G \times H$ into H satisfying

$$eh = h, \qquad (g_1 g_2)h = g_1(g_2 h) \quad \forall h \in H \forall g_1, g_2 \in G,$$

where e is the identity of G.

Definition 3.1. *Suppose $\{H_\gamma\}_{\gamma\in\Gamma}$ is a partition of a set H equipped with a induced strict partial order, G is a monoid acting on H. We shall say that H is a stratified G-set if for all $g \in G$ it satisfies the following conditions :*

$$\begin{aligned}
&1) \quad \forall \gamma \, \forall h_1 \, \forall h_2 \, \exists \gamma' (h_1, h_2 \in H_\gamma \Longrightarrow gh_1, gh_2 \in H_{\gamma'});\\
&2) \quad h_1 \prec h_2 \Longrightarrow gh_1 \prec gh_2;\\
&3) \quad h \prec gh \qquad \forall h \in H \, \forall g \in G \quad (g \neq e),
\end{aligned}$$

where e is the identity of G.

Remark. The above definition is a generalization of ranking [6].

Define an action of the monoid $(\mathbb{N}^n, +)$ on the set of coordinate function U by the rule

$$\beta u^i_\alpha = u^i_{\alpha+\beta} \qquad \forall \alpha, \beta \in \mathbb{N}^n \forall i \in \mathbb{M}$$

with $\mathbb{M} = \mathbb{N}_m$ or $\mathbb{M} = \mathbb{N}$. It is easy to see that $U = \bigcup_{n\in\mathbb{N}} U_n$ with $U_n = \{u^i_\alpha \in U : |\alpha| = n\}$ gives an example of stratified $\mathbb{N}^n$-set.

Let V be an open set in $\mathbb{J}$ and X is given by (2.2). We consider two sets

$$\mathcal{F}(V)_X = \{f \in \mathcal{F}(V) : ivf \subset X\}, \qquad \hat{\mathcal{F}}(V) = \mathcal{F}(V) \setminus \mathcal{F}(V)_X. \tag{3.3}$$

We shall indicate how a partition $\{U_\gamma\}_{\gamma\in\Gamma}$ of the set U leads to a partition of $\hat{\mathcal{F}}(V)$.

Consider sets

$$Y_\gamma = X \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma' \leqslant \gamma} U_{\gamma'} \Big), \qquad \gamma_0 = \min\{\gamma \in \Gamma\} \tag{3.4}$$

which form an ascending chain of subsets of Y. The sets

$$J^\gamma = \{z \in \mathbb{J} : y(z) = 0, \forall y \in (Y \setminus Y_\gamma)\}, \tag{3.5}$$

$$\mathcal{F}^\gamma(V) = \{f \in \mathcal{F}(V) : iv(f) \subset Y_\gamma\} \tag{3.6}$$

also form ascending chains of subspaces and subalgebras respectively. This chain of subalgebras generates a partition $\{\Phi^\gamma(V)\}_{\gamma\in\Gamma}$ of the set $\hat{\mathcal{F}}(V)$, where

$$\Phi^\gamma(V) = \mathcal{F}^\gamma(V) \setminus \Big(\bigcup_{\gamma_0 < \gamma' < \gamma} \mathcal{F}^{\gamma'}(V) \Big), \qquad \Phi^{\gamma_0}(V) = \mathcal{F}^{\gamma_0}(V) \setminus \mathcal{F}(V)_X. \tag{3.7}$$

Let us consider three set of germs

$$\mathcal{F}_a^\gamma = \{\tilde{f} \in \mathcal{F}_a : iv(\tilde{f}) \subset \tilde{Y}_\gamma\}, \tag{3.8}$$

$$\mathcal{F}_{aX} = \{\tilde{f} \in \mathcal{F}_a : iv\tilde{f} \subset \tilde{X}\}, \qquad \hat{\mathcal{F}}_a = \mathcal{F}_a \setminus \mathcal{F}_{aX}. \tag{3.9}$$

A partition $\{\Phi_a^\gamma\}_{\gamma\in\Gamma}$ of the set $\hat{\mathcal{F}}_a$ is given by

$$\Phi_a^\gamma = \mathcal{F}_a^\gamma \setminus (\bigcup_{\gamma_0 < \gamma' < \gamma} \mathcal{F}_a^{\gamma'}), \qquad \Phi_a^{\gamma_0} = \mathcal{F}_a^{\gamma_0} \setminus \mathcal{F}_{aX}. \tag{3.10}$$

Lemma 3.2. *Suppose that the set U (2.2) is a stratified $\mathbb{N}^n$-set. Then the sets $\hat{\mathcal{F}}(V)$ and $\hat{\mathcal{F}}_a$ are also stratified $\mathbb{N}^n$-sets.*

Proof. It suffices to check three requirements of a stratified set for generators of the monoid $\mathbb{N}^n$. At first, we consider the set $\hat{\mathcal{F}}(V)$. To prove first property of a stratified set it will suffice to show the following statement. If $f_1, f_2 \in \Phi^\gamma(V)$, then there exists an element $\gamma' \in \Gamma$ such that functions $\mathcal{D}_k(f_1), \mathcal{D}_k(f_2)$ given by (2.3) lie in $\Phi^{\gamma'}(V)$. We remark that if $\dfrac{\partial f}{\partial u_\alpha^i}$ vanishes on some open set Ω in $\mathbb{J}$ then the function f does not depend on u_α^i in Ω. Since $f_1, f_2 \in \Phi^\gamma(V)$, then there are variables $u_\alpha^i, u_\beta^j \in U^\gamma$, and points $a_1, a_2 \in V$ such that

$$\frac{\partial f_1}{\partial u_\alpha^i}(a_1) \neq 0, \qquad \frac{\partial f_2}{\partial u_\beta^j}(a_2) \neq 0.$$

It follows from assumption of our Lemma that for all $u_\alpha^i, u_\beta^j \in U^\gamma$ there exists $\gamma' \in \Gamma$ such that $\mathcal{D}_k u_\alpha^i, \mathcal{D}_k u_\beta^j$ lie in $U^{\gamma'}$. Thus, we clearly obtain $\dfrac{\partial f_1}{\partial u_\alpha^i} u_{\alpha+e_k}^i, \dfrac{\partial f_2}{\partial u_\beta^j} u_{\beta+e_k}^j \in \Phi^{\gamma'}(V)$ and furthermore, $\mathcal{D}_k f_1, \mathcal{D}_k f_2 \in \Phi^{\gamma'}(V)$. In a similar manner, one can prove two other properties.

We shall now prove that $\hat{\mathcal{F}}_a$ is also a stratified $\mathbb{N}^n$-set. At first, we show that if $\tilde{f} \in \Phi_a^\gamma$ then for any representative f of the germ $\tilde{f}$ there exists a an neighborhood $V^\star$ of a such that for every neighborhood $V' \subset V^\star$ of a there are a variable $u_\alpha^i \in U^\gamma$ and a point $b \in V'$ with $\dfrac{\partial f}{\partial u_\alpha^i}(b) \neq 0$. Suppose this is not the case. Then there exists a representative f of the germ $\tilde{f}$ such that for every neighborhood $V^\star$ of a there is a neighborhood $V' \subset V^\star$ of a in which $\dfrac{\partial f}{\partial u_\alpha^i}(b) = 0$, for any variable $u_\alpha^i \in U^\gamma$ and every point $b \in V'$. Therefore, the function f does not depend on

variables $u^i_\alpha \in U^\gamma$ in neighborhood V'. We thus get a contradiction to $\tilde f \in \Phi^\gamma_a$. This implies that $\dfrac{\partial \tilde f}{\partial \tilde u^i_\alpha} \neq 0$.

Let us prove the first property of a stratified set for $\hat{\mathcal{F}}_a$. Suppose that $\tilde f_1$ and $\tilde f_2$ lie in Φ^γ_a. It suffices to show that $\mathcal{D}_k\tilde f_1$ and $\mathcal{D}_k\tilde f_2$ lie in $\Phi^{\gamma'}_a$ for some $\gamma' \in \Gamma$. From assumption of this lemma, there is an element $\gamma' \in \Gamma$ such that

$$\mathcal{D}_k\tilde u^i_\alpha = \tilde u^i_{\alpha+e_k}, \quad \mathcal{D}_k\tilde u^j_\beta = \tilde u^j_{\beta+e_k} \quad \forall \tilde u^i_\alpha, \tilde u^j_\beta \in \tilde U^\gamma.$$

It follows as above that there are variables $u^i_\alpha, u^j_\beta \in U^\gamma$, an element $\gamma' \in \Gamma$ and a number $k \in \mathbb{N}$ such that germs $\dfrac{\partial \tilde f_1}{\partial \tilde u^i_\alpha}\tilde u^i_{\alpha+e_k}, \dfrac{\partial \tilde f_2}{\partial \tilde u^j_\beta}\tilde u^j_{\beta+e_k}$ lie in $\Phi^{\gamma'}_a$. Hence $\mathcal{D}_k\tilde f_1, \mathcal{D}_k\tilde f_2 \in \Phi^{\gamma'}_a$. The other properties are proved in the same vein.

In what follows we shall suppose that $\hat{\mathcal{F}}(V)$ and $\hat{\mathcal{F}}_a$ are stratified $\mathbb{N}^n$-sets equipped with a induced strict partial order. □

Definition 3.3.

(i) A function $f = u^i_\alpha + g \in \mathcal{F}(V)$ is called orderly solvable (with respect to u^i_α), if $g \prec u^i_\alpha$. The variable u^i_α is denoted by ltf and is called leading term of f.

(ii) A germ $\tilde f = \tilde u^i_\alpha + \tilde g \in \mathcal{F}_a$ is called orderly solvable (with respect to $\tilde u^i_\alpha$) if $\tilde g \prec \tilde u^i_\alpha$. The germ $\tilde u^i_\alpha$ is denoted by $lt\tilde f$ and is called leading term of $\tilde f$.

Proposition 3.4. *Let $\tilde F \in \mathcal{F}_a$ be a germ depending on $\tilde u^i_\beta$. Suppose that $\tilde f = \tilde u^i_\alpha + \tilde g$ is a orderly solvable germ with respect to $\tilde u^i_\alpha$ and there exists $\delta \in \mathbb{N}^n$ satisfying $\beta = \alpha + \delta$. Then there exists a unique germ $\tilde r \in \mathcal{F}_a$ and a germ $q \in \mathcal{F}_a$ such that*

$$\tilde F = \tilde q D^\delta \tilde f + \tilde r, \quad \tilde u^i_\alpha \notin iv\tilde r \tag{3.11}$$

$$\tilde q \preceq \tilde F, \quad \tilde r \preceq \tilde F. \tag{3.12}$$

Proof. The germ $D^\delta\tilde f$ is equal to $\tilde u^i_\beta + D^\delta\tilde g$, where $D^\delta\tilde g \prec \tilde u^i_\beta$. Then from the Mather theorem [15], we obtain (3.11). The uniqueness $\tilde r$ is proved just as in the second part of Proposition 1.3. It is clear that

$$iv(\tilde q) \subseteq (iv(\tilde F) \cup iv(D^\delta\tilde g)), \quad iv(\tilde r) \subset (iv(\tilde F) \cup iv(D^\delta\tilde g)), \quad \tilde{u^i_\alpha} \notin iv\tilde r.$$

Since $lt(D^\delta\tilde g) = \tilde u^i_\alpha$, it follows that $D^\delta\tilde g \preceq \tilde F$. The last relations lead to (3.12). □

If the assumptions of Proposition 3.4 are satisfied, then we say that the *the germ $\tilde F$ reduces to $\tilde r$ modulo $\tilde f$* at a, denoted by $\tilde F \underset{\tilde f}{\to} \tilde r$.

Proposition 3.5. *Let F be a polynomial in u^i_β with coefficients that lie in $\mathcal{F}(V)$ and do not depend on u^i_β. Assume that $f = u^i_\alpha + g$ is a orderly solvable function with respect to u^i_α and δ is a element in $\mathbb{N}^n$ satisfying $\beta = \alpha + \delta$. Then there exists a unique function $\tilde r \in \mathcal{F}(V)$ and a function $q \in \mathcal{F}(V)$ such that*

$$F = qD^\delta f + r, \quad u^i_\alpha \notin ivr \tag{3.13}$$

$$q \preceq F, \quad r \preceq F. \tag{3.14}$$

Proof. The relation (3.13) follows from Proposition 1.4. The inequalities (3.14) are proved just as in the second part of Proposition 3.4. □

If the assumptions of Proposition 3.5 are satisfied, then we say that the *the function F reduces to the function r modulo f* on V, denoted by $F \underset{f}{\to} r$.

Definition 3.6. *A differential system $S \subset \mathcal{F}(V)$ is called weakly solvable if every function $f \in S$ is orderly solvable. We write $ltS = \{ltf : f \in S\}$.*

It is clear that if a germ $\tilde{f} \in \mathcal{F}_a$ is orderly solvable with respect to $\tilde{u}^i_\alpha$, then it is solvable with respect to $\tilde{u}^i_\alpha$ in terms of Definition 2.5. In the future, we suppose that $st\tilde{f} = lt\tilde{f}$ in this case. Furthermore, we assume that $st\tilde{S} = lt\tilde{S}$ for every weakly solvable system.

Definition 3.7. *Let $S \subset \mathcal{F}(V)$ be a weakly solvable differential system.*

(i) Let a be a point in V. We shall say that a germ $\tilde{F} \in \mathcal{F}_a$ reduces to a germ $\tilde{r} \in \mathcal{F}_a$ modulo $\tilde{S}_a$, written $\tilde{F} \underset{\tilde{S}}{\longrightarrow} \tilde{r}\big|_a$, if there exists a consequence of germs $\tilde{r}_1, \dots, \tilde{r}_{k-1} \in \tilde{\mathcal{F}}_a$ such that

$$\tilde{F} \underset{\tilde{f}_1}{\longrightarrow} \tilde{r}_1 \underset{\tilde{f}_2}{\longrightarrow} \cdots \underset{\tilde{f}_{k-1}}{\longrightarrow} \tilde{r}_{k-1} \underset{\tilde{f}_k}{\longrightarrow} \tilde{r}$$

with $\tilde{f}_1, \dots, \tilde{f}_k \in \tilde{S}_a$.

(ii) Let S be a normalized set in $\mathcal{F}(V)$. Suppose that $F \in \mathcal{F}(V)$ is a polynomial in $O(ltS)$ with coefficients being in $\mathcal{F}(V)$ and depending only on variables in $O(Y \setminus ltS)$. We say that F reduces to a function $r \in \mathcal{F}(V)$ modulo S, written $F \underset{S}{\longrightarrow} r$ if there exists a consequence of functions $r_1, \dots, \tilde{r}_{k-1} \in \mathcal{F}(V)$ such that

$$F \underset{f_1}{\longrightarrow} r_1 \underset{f_2}{\longrightarrow} \cdots \underset{f_{k-1}}{\longrightarrow} r_{k-1} \underset{f_k}{\longrightarrow} r$$

with $f_1, \dots, f_k \in S$.

Let us define a binary operation $\diamond$ on $\mathbb{N}^n$ by

$$\alpha \diamond \beta = (\mu_1, \dots, \mu_n),$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$, $\beta = (\beta_1, \dots, \beta_n)$, $\mu_i = \max(\alpha_i, \beta_i) - \alpha_i$. Suppose that functions $f_1, f_2 \in \mathcal{F}(V)$ are orderly solvable with respect to u^i_α, u^i_β respectively and $\tilde{f}_1, \tilde{f}_2$ are their germs at $a \in V$. Then we define two differences

$$\tau(f_1, f_2) = D^{\alpha\diamond\beta} f_1 - D^{\beta\diamond\alpha} f_2, \qquad \tau(\tilde{f}_1, \tilde{f}_2) = D^{\alpha\diamond\beta} \tilde{f}_1 - D^{\beta\diamond\alpha} \tilde{f}_2. \tag{3.15}$$

Definition 3.8. *Let $S \subset \mathcal{F}(V)$ be a weakly solvable differential system.*

(i) The system S satisfies reducibility conditions at $a \in V$ if

$$\tau(\tilde{f}_1, \tilde{f}_2) \underset{\tilde{S}}{\longrightarrow} \tilde{0}\big|_a \tag{3.16}$$

for each pair of functions $f_1, f_2 \in S$ such that $ltf_1 = u^i_\alpha$, $ltf_2 = u^i_\beta$.

(ii) Let S be a normalized set in $\mathcal{F}(V)$. We say that S satisfies reducibility conditions on V if

$$\tau(f_1, f_2) \underset{S}{\longrightarrow} 0 \tag{3.17}$$

for each pair of functions $f_1, f_2 \in S$ such that $ltf_1 = u^i_\alpha$, $ltf_2 = u^i_\beta$.

Denote by $\mathbb{D}$ an algebra of operators such that every element of $\mathbb{D}$ can be written as a finite sum

$$P = \sum a_\alpha D^\alpha \tag{3.18}$$

with $a_\alpha \in \mathbb{R}$. Let RU be a vector space over $\mathbb{R}$ consisting of finite sums

$$s = \sum b^\beta_i u^i_\beta, \qquad b^\beta_i \in \mathbb{R}. \tag{3.19}$$

Define an action of $\mathbb{D}$ on RU by letting

$$P u^i_\beta = \sum a_\alpha u^i_{\alpha+\beta},$$

and extending P to RU by linearity.

Definition 3.9. *Let* $\mathbf{y}$ *be an* k*-tuple* $(y_{t_1},\dots,y_{t_k})$ *of variables* $y_{t_i}\in U$. *An* k*-tuple* $\mathbf{d}=(d_1,\dots,d_k)$ *of operators in* $\mathbb{D}$ *is called syzygy of* $\mathbf{y}$, *if*

$$d_1y_{t_1}+\cdots+d_ky_{t_k}=0.$$

The syzygies of the k*-tuple* $\mathbf{y}$ *constitute a* $\mathbb{D}$*-module denoted by* $Syz\ \mathbf{y}$.

Suppose $\mathbf{y}=(y_{t_1},\dots,y_{t_k})\in U^k$ with $y_{t_i}=u^l_\alpha$ and $y_{t_j}=u^l_\beta$, then

$$\sigma_{ij}=\mathcal{D}^{\alpha\diamond\beta}e_i-\mathcal{D}^{\beta\diamond\alpha}e_j \tag{3.20}$$

is a syzygy of $\mathbf{y}$. It is easy to show (see [11]) that the syzygies (3.20) generate the $\mathbb{D}$-module $Syz\ \mathbf{y}$ if a number of the indeterminates $u^i_0\in U$ is finite.

Example. Assume $m=1$ and $n=2$, so that $U=\{u_{(i,j)}: i,j\in\mathbb{N}\}$; take $\mathbf{y}=(u_{(0,1)},u_{(0,2)},u_{(1,1)})$. It is obvious that $(\mathcal{D}_2,-1,0)$, $(\mathcal{D}_1,0,-1)$ and $(0,\mathcal{D}_1,-\mathcal{D}_2)$ are syzygies of the 3-tuple $\mathbf{y}$.

4. Passivity criterion of differential systems

In this section we give sufficient conditions for a differential system to be passive. Furthermore, we prove that any passive system generates a manifold in the jet space.

Let $S\subset\mathcal{F}(V)$ be a weakly solvable differential system. We call a point $a\in\mathbb{J}$ equivalent to a point $b\in\mathbb{J}$, written $a\sim b$, if $y(a)=y(b)$ for all coordinate functions $y\in Y\setminus O(ltS)$.

Theorem 4.1. *Let* $S=\{f_1,\dots,f_k\}\subset\mathcal{F}(V)$ *be a differential system with finite number of indeterminates. Suppose that* S *is a normalized set and satisfies reducibility conditions (3.16) at* $a\in V$. *Then the following properties hold:*

(1) there is a unique point $b\sim a$ *such that*

$$D^\alpha f(b)=0,\quad \forall f\in S\ \forall\alpha\in\mathbb{N}^n; \tag{4.1}$$

(2) the system S *is passive at any point* $c\sim a$.

Proof. Since S is a normalized set, we conclude that the orbit $O(S)$ is a weakly solvable differential system. This gives rise to the uniqueness of the point b satisfying the condition (4.1).

We have shown above that a partition $\{U_\gamma\}_{\gamma\in\Gamma}$ of the set U provides the ascending chain of subspaces J^γ (3.5), the chains of subalgebras $\mathcal{F}^\gamma(V)$ (3.6), $\mathcal{F}^\gamma_z$ (3.8) and leads to the partitions $\{\Phi^\gamma(V)\}_{\gamma\in\Gamma}$ (3.7), $\{\Phi^\gamma_z\}_{\gamma\in\Gamma}$ (3.10) with $z\in V$. We also recall that Y_γ is defined by (3.4). Consider linear mappings $\pi_\gamma:\mathbb{J}\longrightarrow J^\gamma$, where the coordinates of $\pi_\gamma(z)$ are given by

$$y(\pi_\gamma(z))=y(z)\quad\forall y\in Y_\gamma;\qquad y(\pi_\gamma(z))=0\quad\forall y\in Y\setminus Y_\gamma\ .$$

Recall that $\tilde S_z$ is a set of germs of functions in S at the point z. We shall use the following notion:

$$\gamma_0=\min\{\gamma\in\Gamma: O(\tilde S_a)\cap\Phi^\gamma_a\neq 0\},$$
$$O^\gamma_z=O(\tilde S_z)\cap\mathcal{F}^\gamma_z,\quad C^\gamma_z=O(\tilde S_z)\cap\Phi^\gamma_z.$$

It is obvious that

$$O^{\gamma_\star}_z=C^{\gamma_\star}_z\cup\Big(\bigcup_{\gamma_0\leqslant\gamma<\gamma_\star}C^\gamma_z\Big) \tag{4.2}$$

for any $\gamma_\star>\gamma_0$. Let $\langle O^\gamma_z\rangle$ be an ideal of the algebra $\mathcal{F}^\gamma_z$ generated by O^γ_z.

We shall use transfinite induction to prove that for all $\gamma\geqslant\gamma_0$ the following properties hold:

(i) there exists a point $b_\gamma\sim\pi_\gamma(a)$ such that $\tilde f(b_\gamma)=0$ for all $\tilde f\in O^\gamma_{b_\gamma}$;

(ii) there exists a normalized system $\tilde{B}_c^{\gamma}$ of generators of the ideal $\langle O_c^{\gamma}\rangle$ for any point $c \sim \pi_{\gamma}(a)$.

Assume that $\gamma = \gamma_0$ then two cases arise:

1. All leading terms of germs in $C_a^{\gamma_0}$ are distinct.
2. There exist at least two germs $\tilde{f}_i, \tilde{f}_j \in C_a^{\gamma_0}$ such that $lt\tilde{f}_i = lt\tilde{f}_j$.

It is clear that in the first case there is a point $b_{\gamma_0} \sim \pi_{\gamma_0}(a)$ such that $\tilde{f}(b_{\gamma_0}) = 0$ for all $\tilde{f} \in C_{b_{\gamma_0}}^{\gamma_0}$. Furthermore, the properties (ii) and (iii) are satisfied because $\tilde{B}_c^{\gamma_0} = \tilde{C}_c^{\gamma_0}$ is a normalized system of generators of the ideal $\langle O_c^{\gamma_0}\rangle$ and $B^{\gamma_0} = S^{\gamma_0}$ is a normalized system of generators of the ideal $\langle O_c^{\gamma_0}\rangle$. In the second case, there must be germs $\tilde{f}_i, \tilde{f}_j \in \tilde{C}_a^{\gamma_0}$ such that $lt\tilde{f}_i = lt\tilde{f}_j$. Then $\tilde{f}_i - \tilde{f}_j \in \mathcal{F}_a^{\gamma'}$, where $\gamma' < \gamma_0$, and $\tilde{f}_i - \tilde{f}_j \underset{\tilde{S}_a}{\longrightarrow} \tilde{0}$ according to the conditions of our theorem. Since $O_a^{\gamma'}$ is the empty set then we have $\tilde{f}_i = \tilde{f}_j$.

Assume that our statement is true for all γ with $\gamma_0 \leqslant \gamma < \gamma_\star$ and prove its for $\gamma = \gamma_\star$. As above, we need to distinguish two cases:

1. All leading terms of germs in $C_a^{\gamma_\star}$ are distinct.
2. There exist two germs $\tilde{f}, \tilde{g} \in C_a^{\gamma_\star}$ such that $lt\tilde{f} = lt\tilde{g}$.

In the first case, the property (i) is trivially satisfied. According to the assumption of induction and the formula (4.2), the set

$$G_c^{\gamma_\star} = C_c^{\gamma_\star} \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma < \gamma_\star} \tilde{B}_c^{\gamma} \Big)$$

is a system of generators (not necessarily normalized) of the ideal $\langle O_c^{\gamma_\star}\rangle$ for any point $c \sim \pi_{\gamma_\star}(a)$.

In the second case, there are two germs $\tilde{f}, \tilde{g} \in C_a^{\gamma_\star}$ with $lt\tilde{f} = lt\tilde{g}$. Then there exist two germs $\tilde{f}_p, \tilde{f}_q \in \tilde{S}_a$ such that

$$lt\tilde{f} = ltD^{\mu}\tilde{f}_p = lt\tilde{g} = ltD^{\eta}\tilde{f}_q,$$

where $D^{\mu} = D_1^{\mu_1} \cdots D_n^{\mu_n}$ and $D^{\eta} = D_1^{\eta_1} \cdots D_n^{\eta_n}$ are some differential monomials. Therefore, we have

$$D^{\mu}(lt\tilde{f}_p) = D^{\eta}(lt\tilde{f}_q). \tag{4.3}$$

Denote by $\mathbf{y}$ an n-tuple constructed from all elements of the set $lt\tilde{S}_a$. Assume that the elements $lt\tilde{f}_p$ and $lt\tilde{f}_q$ are the i-th and j-th items in $\mathbf{y}$. It follows from (4.3) that $d = D^{\mu}e_i - D^{\eta}e_j$ is a syzygy of $\mathbf{y}$. It is easy to see that there is differential monomial D^{ν} such that $d = D^{\nu}\sigma_{ij}$, where σ_{ij} is one of the sygyzies (3.20) generating $\mathbb{D}$-module $Syz\ \mathbf{y}$.

The difference $\tilde{f} - \tilde{g}$ reduces to the zero germ modulo $\tilde{S}_a$. Indeed, the system S satisfies reducibility conditions at a by assumption, then we have

$$D^{\nu}\sigma_{ij}(\tilde{f}_p, \tilde{f}_q) = \tilde{f} - \tilde{g} \underset{\tilde{S}_a}{\longrightarrow} \tilde{0}.$$

Next, we include any one of the germs $\tilde{f}, \tilde{g}$ in a new set $gen_a^{\gamma_\star}$ while the other is not. In the same way we inspect all pairs of germs in $C_a^{\gamma_\star}$ with equal leading terms, form the set $gen_a^{\gamma_\star}$ and obtain a system of generators

$$G_a^{\gamma_\star} = gen_a^{\gamma_\star} \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma < \gamma_\star} \tilde{B}_a^{\gamma} \Big)$$

for the ideal $\langle O_a^{\gamma_\star}\rangle$.

We now prove the existence of a normalized system of generators for ideal $\langle O_a^{\gamma_\star}\rangle$. Any germ $f \in G_a^{\gamma_\star} \cap \Phi_a^{\gamma_\star}$ is of the form $\tilde{f} = \tilde{u}_{\alpha}^{i} + \tilde{h}$ with $\tilde{h} \in \mathcal{F}_a^{\gamma}$ and $\gamma < \gamma^\star$. According to Proposition 1.3 and the assumption step of induction, the germ $\tilde{h}$ is represented by

$$\tilde{h} = \tilde{q_1}\tilde{f}_{t_1} + \cdots + \tilde{q_p}\tilde{f}_{t_p} + \tilde{r},$$

where $\tilde{f}_{t_i} \in \tilde{B}_a^\gamma$, $\tilde{q}_i \in \mathcal{F}_a^\gamma$, and the germ $\tilde{r} \in \mathcal{F}_a^\gamma$ does not depend on principal variables of $\tilde{B}_a^\gamma$. Then the germ $\tilde{f}^\star = \tilde{u}_\alpha^i + \tilde{r}$ is included in a set $ben_a^{\gamma\star}$. To do so with every germ in $G_a^{\gamma\star} \cap \Phi_a^{\gamma\star}$, we obtain an normalized system of generators

$$\tilde{B}_a^{\gamma\star} = ben_a^{\gamma\star} \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma < \gamma_\star} \tilde{B}_a^\gamma \Big)$$

for the ideal $\langle O_a^{\gamma\star} \rangle$.

Let us take a point $c \sim \pi_{\gamma_\star}(a)$, then the ideal $\langle O_a^{\gamma\star} \rangle$ is isomorphic to the ideal $\langle O_c^{\gamma\star} \rangle$ of the algebra $\mathcal{F}_a^{\gamma\star}$. Indeed, if a function f lies in S, then $\tilde{f}_a = \tilde{u}_\alpha^i|_a + \tilde{g}_a$ and $\tilde{f}_c = \tilde{u}_\alpha^i|_c + \tilde{g}_c$ because S is a normalized set. Since $c \sim \pi_{\gamma_\star}(a)$, then $\tilde{g}_a = \tilde{g}_c$ and the ideal $\langle O_a^{\gamma\star} \rangle$ is isomorphic to the ideal $\langle O_c^{\gamma\star} \rangle$. Therefore, the ideal $\langle O_c^{\gamma\star} \rangle$ has a normalized system of generators.

It is easy to show that a point b, such that $\pi_\gamma(b) = b_\gamma$ for all $\gamma \in \Gamma$, satisfies (4.1) and the set

$$\tilde{B}_c = \bigcup_{\gamma_0 \leqslant \gamma} \tilde{B}_c^\gamma$$

is a normalized system of generators for the differential ideal $\langle\langle \tilde{S} \rangle\rangle_c$ of $\mathcal{F}_c$. Therefore, the ideal $\langle\langle \tilde{S} \rangle\rangle_c$ is soft. By construction, we see that the set $\tilde{B}_c$ coincides with the orbit $O(lt\tilde{S}_c)$. Thus S is a passive system at $c \sim a$ and the theorem is proved. □

Theorem 4.2. *Let $S = \{f_1, \dots, f_k\} \subset \mathcal{F}(V)$ be a differential system with finite number of indeterminates. Suppose that S is a normalized set and satisfies reducibility conditions (3.17) on V. Then the system S is passive on V and the set*

$$\mathcal{M} = \{z \in V : f(z) = 0, f \in O(S)\} \tag{4.4}$$

is a manifold in $\mathbb{J}$.

Our proof is almost the same as the proof of Theorem 4.1. We employ the following denotation:

$$\gamma_0 = \min\{\gamma \in \Gamma : O(S) \cap \Phi^\gamma(V) \neq 0\}, \quad O^\gamma = O(S) \cap \mathcal{F}^\gamma(V),$$
$$C^\gamma = O(S) \cap \Phi^\gamma(V), \quad S^\gamma = O(S) \cap \mathcal{F}^\gamma(V).$$

Let $\langle O^\gamma \rangle$ be an ideal of the algebra $\mathcal{F}^\gamma(V)$ generated by O^γ.

Using transfinite induction, we prove that for all $\gamma \geqslant \gamma_0$ there exists a normalized system of generators of the ideal $\langle O^\gamma \rangle$. Just as in the above theorem, we see that O^{γ_0} is a normalized system of generators of the ideal $\langle O^{\gamma_0} \rangle$.

Suppose that for each $\gamma_0 \leqslant \gamma < \gamma_\star$ there exists a normalized system of generators B^γ of the ideal $\langle O^\gamma \rangle$. We need to check the existences of such a system for $\gamma = \gamma_\star$. At first one obtain a special system of generators $G^{\gamma_\star}$ of the ideal $\langle O^{\gamma_\star} \rangle$. For this purpose, we consider the two cases again:

(1) All leading terms of functions in $C^{\gamma_\star}$ are distinct.

(2) There exist at least two functions $f_i, f_j \in C^{\gamma_\star}$ such that $ltf_i = ltf_j$.

In the first case the set

$$G^{\gamma_\star} = C^{\gamma_\star} \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma < \gamma_\star} B^\gamma \Big)$$

is a system of generators of the ideal $\langle O^{\gamma_\star} \rangle$ In the second case there exist functions $f, g \in C^{\gamma_\star}$ such that $ltf = ltg$. Thus there exist functions $f_p, f_q \in S$ and elements $\mu, \nu \in \mathbb{N}^n$ satisfying

$$ltf = D^\mu(ltf_p) = D^\nu ltf_q = ltg.$$

It follows from condition of our theorem that the difference $f-g$ reduces to the zero function modulo S. One of the functions f, g is included in a set $gen^{\gamma_\star}$. In the same way we search for all pairs of functions in $C^{\gamma_\star}$ with equal leading terms, form the set $gen^{\gamma_\star}$ and obtain a system of generators

$$G^{\gamma_\star} = gen^{\gamma_\star} \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma < \gamma_\star} B^{\gamma} \Big)$$

for the ideal $\langle O^{\gamma_\star} \rangle$.

We can then construct the set $B^{\gamma_\star}$ as follows. Any function $f \in G^{\gamma_\star}$ is the form $u^i_\alpha + h$, where $h \in \mathcal{F}^{\gamma}(V)$ with $\gamma < \gamma_\star$. Furthermore the function h is a polynomial in principal variables of B^{γ} and coefficients of this polynomial depend only on parametric variables.

Using Preposition 1.4, we write

$$h = \sum q_i f_{t_i} + r,$$

where $f_{t_i} \in B^{\gamma}$, $q_i \in \mathcal{F}^{\gamma}(V)$, and the function $r \in \mathcal{F}^{\gamma}(V)$ depends only on parametric variables. We include then the function $f^\star = u^i_\alpha + r$ in a set $ben^{\gamma_\star}$. To do so with every function in $G^{\gamma_\star} \cap \Phi^{\gamma_\star}$, we obtain an normalized system of generators

$$B^{\gamma_\star} = ben^{\gamma_\star} \cup \Big(\bigcup_{\gamma_0 \leqslant \gamma < \gamma_\star} B^{\gamma} \Big)$$

for the ideal $\langle O^{\gamma_\star} \rangle$. The set

$$B = \bigcup_{\gamma_0 \leqslant \gamma} B^{\gamma}$$

is a normalized system of generators for the ideal $\langle\langle S \rangle\rangle$ of $\mathcal{F}(V)$ and this ideal is soft. It is easy to see that the set B coincides with the orbit $O(ltS)$. Thus the system S is a passive on V. From Proposition 1.7 it follows that the set (4.4) is a manifold in $\mathbb{J}$ and the theorem is proved.

5. Examples

We exhibit some examples assuming that $n=2$, $m=1$ and denoting by u the variable u_{00}. The sets $U_n = \{u_{ij} : i+j=n\}$ form a partition of $U = \{u_{ij}\}_{i,j\in\mathbb{N}}$. We shall sometimes apply the usual terminology of differential equations.

The smooth function

$$f = u_{11} - \sinh u \tag{5.1}$$

corresponds to the partial differential equation

$$u_{tx} - \sinh u = 0. \tag{5.2}$$

It is known (see [19]) that vector fields

$$X_1 = (u_{03} - \frac{1}{2}u_{01}^3)\frac{\partial}{\partial u} + \cdots, \quad X_2 = (u_{05} - \frac{5}{2}u_{01}^2 u_{03} - \frac{5}{2}u_{01}u_{02}^2 + \frac{3}{8}u_{01}^5)\frac{\partial}{\partial u} + \cdots$$

are higher symmetries of the equation (5.2).

Let S_1 be a differential system consisting of the functions f and $h_1 = u_{03} - \frac{1}{2}u_{01}^3$. We want to show that the ideal $I = \langle\langle S_1 \rangle\rangle$ is soft. For this purpose we shall construct a passive system generating the ideal I. The functions f and h_1 are orderly solvable with respect to u_{11} and u_{03} respectively. It is a straightforward calculation to check that the function $\tau(f, h_1)$ (given by (3.15)) reduces to the function $f_1 = u_{02} - \frac{1}{2}u_{01}^2 \tanh(u)$ modulo S_1. Then an easy calculation

shows that the function $\tau(f, f_1)$ reduces to the function $f_2 = u_{10} - 2\cosh(u)/u_{01}$ modulo f. It is easy to see that the function $\tau(f_1, f_2)$ reduces to 0 modulo the system $S = \{f_1, f_2\}$. Furthermore the system S generates the ideal I and is passive.

We now find solutions of the system S. The function f_1 produces the ordinary differential equation

$$u_{xx} - \frac{1}{2}u_x^2 \tanh(u) = 0$$

having the first integral $u_x/\cosh^2 u$. Using this integral and the equation

$$u_t = 2\frac{\cosh u}{u_x},$$

we obtain the implicit solution

$$\int \frac{du}{\sqrt{\cosh u}} = cx - 2t/c + c_1$$

with $c, c_1 \in \mathbb{R}$.

Consider now the system S_2 consisting of the functions f and $h_2 = u_{05} - \frac{5}{2}u_{01}^2 u_{03} - \frac{5}{2}u_{01}u_{02}^2 + \frac{3}{8}u_{01}^5$. A direct calculation shows that the function $\tau(f, h_2)$ reduces to the function

$$f_3 = u_{04} - u_{01}u_{03}\tanh u + \frac{1}{2}u_{02}^2 \tanh u - \frac{3}{2}u_{01}^2 u_{02} + \frac{3}{8}u_{01}^4 \tanh u$$

modulo S_2. Then the function $\tau(f, f_3)$ reduces to

$$f_4 = u_{10} + \frac{4(u_{01}^3 - 2u_{03})\cosh u}{8u_{01}u_{03} - 4u_{02}^2 - 3u_{01}^4}$$

modulo $S_3 = \{f_3, f_4\}$. It is possible to check that the function $\tau(f_3, f_4)$ reduces to 0 modulo $S_4 = \{f_3, f_4\}$, the system S_4 is passive and it generates the soft ideal $\langle\langle S_2 \rangle\rangle$.

The next example is closely connected with the equation (5.2) as well. The set

$$v_{xxx} - \frac{2v_x v_{xx}}{v} + rv^4 + s = 0, \qquad r, s \in \mathbb{R},$$

is invariant manifold of the partial differential equation

$$v_t = v_{xx}/v$$

as shown in [20]. Using the transformation $v = \exp w$ we rewrite the last equations as

$$w_{xxx} + w_x w_{xx} - w_x^3 + r\exp(3w) + s\exp(-w) = 0, \qquad w_t - (w_{xx} + w_x^2)\exp(-w) = 0.$$

These equations correspond to two functions

$$f_5 = u_{03} + u_{01}u_{02} - u_{01}^3 + r\exp(3u) + s\exp(-u), \quad f_6 = u_{10} - (u_{02} + u_{01}^2)\exp(-u).$$

It is easy to check that the system $S_5 = \{f_5, f_6\}$ generates a soft ideal $\langle\langle S_5 \rangle\rangle$ and is passive. The function $D_2(f_6)$ reduces to the function $f_7 = u_{11} + r\exp(2u) + s\exp(-2u)$ modulo f_5. This function lies in ideal $\langle\langle S_5 \rangle\rangle$ and corresponds to equation

$$u_{tx} = \sinh(2u)$$

with $r = -1/2$ and $s = 1/2$.

6. Conclusion

We defined the concept of a passive system of equations using algebraic constructions. It is proved that such systems are manifolds in an infinite-dimensional space of jets. Moreover, the space is equipped with the Tikhonov topology. We prefer to use the term passive system as it is classical and the word involution is used in different senses. The passivity criterion is a generalization of the classical case. However, we do not prove the existence of a solution to the passive system since we are dealing with smooth systems.

This work was financially supported by the Russian Foundation for Basic Research (Grant no. 17-01-00332-a).

References

[1] E. Goursat, A Course in Mathematical Analysis, Vol. II, Part Two. Differential Equations, Dover Publications, New York, 1945.

[2] S.V.Meleshko, Methods for Constructing Exact Solutions of Partial Differential Equations. Mathematical and Analytical Techniques with Applications to Engineering, Springer, 2005.

[3] C.Riquier, Les systémes des équations aux dérivées partielles, Paris, Gauthier-Villars, 1910.

[4] M.Janet, Leçons sur les systémes des équations aux dérivées partielles, Paris, Gauthier-Villars, 1929.

[5] E.Cartan, Les systémes differentiels exterieurs et leurs applications scientifiques, Hermann, 1946.

[6] E.R.Kolchin, Differential algebra and algebraic groups, Acad. Press, New York, 1973.

[7] J.F. Pommaret, Systems of Partial Differential Equations and Lie Pseudogroups, New York, London, Paris: Gordon and Breach, 1978.

[8] A.M.Vinogradov, Cohomological Analysis of Partial Differential Equations and Secondary Calculus, Translations of Math. Monographs 204, Amer. Math. Soc., 2001.

[9] M.V.Kondratieva, A.B.Levin, A.V.Mikhalev, Differential and Difference Dimension Polynomials, Kluwer Academic Publishers Dordrecht, Netherlands, 1999.

[10] W.Seiler, Involution: The Formal Theory of Differential Equations and its Applications, in Computer Algebra, Springer-Verlag, Berlin Heidelberg, 2010.

[11] O.V.Kaptsov, Systems of generators for ideals of algebra of convergent differential series, *Program. Comput. Softw.*, **40**(2014), no. 2, 63–70.

[12] O.V.Kaptsov, Local algebraic analysis of differential systems, *Theoretical and Mathematical Physics*, **183**(2015), no. 3, 737–752.

[13] J.Gathen, J.Gerhard, Modern Computer Algebra, 3rd Edition, Cambridge University Press, 2013.

[14] V.V.Zharinov, Lecture notes on geometrical aspects of partial differential equations, World Scientific, Singapore, 1992.

[15] Th.Bröcker, L.Lander, Differentiable Germs and Catastrophes, Cambridge University Press, 1975.

[16] N.Jacobson, Basic Algebra II: Second Edition, Dover Books on Mathematics, NY, 1985.

[17] J.Ritt, Differental algebra, New York, AMS Colloquium Publications, Vol. 33, 1950.

[18] M.Golubitsky, V.Guillemin, Stable mappings and their singularities, New York, Springer-Verlag, 1973

[19] N.H.Ibragimov, Groups Applied to Mathematical Physics, Riedel, Dordrecht, 1985.

[20] O.V.Kaptsov, Integration Methods for Partial Differential Equations, Fizmatlit, Moscow, 2009 (in Russian).

Идеалы, порожденные дифференциальными уравнениями

Олег В. Капцов
Институт вычислительного моделирования СО РАН
Красноярск, Российская Федерация

Аннотация. В работе предлагается новый алгебраический подход к исследованию совместности дифференциальных уравнений. Этот подход использует методы коммутативной алгебры, алгебраической геометрии и базисов Гребнера. Мы получаем достаточные условия пассивности систем уравнений в частных производных и доказываем, что такие системы порождают многообразия в пространстве струй. Представлены примеры исследования пассивности систем, порожденных симметриями уравнения sinh-Cordon.

Ключевые слова: дифференциальные кольца и идеалы, базис Гребнера, уравнения в частных производных.

DOI: 10.17516/1997-1397-2020-13-2-187-196
УДК 517.55+517.962.26

The Cauchy Problem for Multidimensional Difference Equations in Lattice Cones

Alexander P. Lyapin*
Siberian Federal University
Krasnoyarsk, Russian Federation
Lesosibirsk Pedagogical Institute — branch of SFU
Lesosibirsk, Russian Federation
Sreelatha Chandragiri†
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 21.12.2019, received in revised form 26.01.2020, accepted 03.02.2020

Abstract. We consider a variant of the Cauchy problem for a multidimensional difference equation with constant coefficients, which connected with a lattice path problem in enumerative combinatorial analysis. We obtained a formula in which generating function of the solution to the Cauchy problem is expressed in terms of generating functions of the Cauchy data and a formula expressing solution to the Cauchy problem through its fundamental solution and Cauchy data.

Keywords: difference equation, fundamental solution, generating function, Dyck paths.

Citation: A.P.Lyapin, S.Chandragiri, The Cauchy Problem for Multidimensional Difference Equations in Lattice Cones, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 187-196.
DOI: 10.17516/1997-1397-2020-13-2-187-196.

1. Definitions and main results

On complex valued functions $f : \mathbb{Z}^n \to \mathbb{C}$ we define the shift operator δ_j as follows:

$$\delta_j : f(x_1, \dots, x_j, \dots, x_n) \mapsto f(x_1, \dots, x_j + 1, \dots, x_n)$$

and the polynomial difference operator

$$P(\delta) = \sum_{\omega \in \Omega} c_\omega \delta^\omega,$$

where $\Omega \subset \mathbb{Z}^n$ is a finite set of points of an n-dimensional lattice, $\delta^\omega = \delta_1^{\omega_1} \cdot \ldots \cdot \delta_n^{\omega_n}$ and $c_\omega \in \mathbb{C}$ are the coefficients of the difference operator.

We consider the difference equation

$$P(\delta) f(x) = g(x), \quad x \in X, \tag{1}$$

where $f(x)$ is an unknown function, and $g(x)$ is a function defined on some set $X \subset \mathbb{Z}^n$. Also choose a set $X_0 \subset \mathbb{Z}^n$, the points of which will be called initial (boundary) points.

*aplyapin@sfu-kras.ru https://orcid.org/0000-0002-0149-7587
†srilathasami66@gmail.com

In the general situation we have to solve *the Cauchy problem*: find a function $f(x)$, satisfying equation (1) and coinciding with a given function $\varphi(x)$ of initial data on the set X_0 :

$$f(x) = \varphi(x), \quad x \in X_0. \tag{2}$$

The function $g(x)$ in the right-hand side of (1) and the initial data function $\varphi(x)$ in (2) is called *the Cauchy data* of problem (1)–(2).

Existence and uniqueness of problem (1)–(2) (solvability of the Cauchy problem) depends on all the objects involved in its formulation: the difference operator $P(\delta)$, the set X on which the right part of the equation is given, and the set X_0 on which the initial data $\varphi(x)$ is defined.

In the one-dimensional case two variants of the Cauchy problems are usually considered:

(i) $X = \{x \in \mathbb{Z} : x \geqslant 0\}$ is the set of non-negative integers, $P(\delta) = \sum\limits_{\omega=0}^{m} c_\omega \delta^\omega$, $X_0 = \{0, 1, \ldots, m-1\}$, $c_m \neq 0$,

(ii) $X = \{x \in \mathbb{Z} : x \geqslant m\}$, $P(\delta) = \sum\limits_{\omega=0}^{m} c_\omega \delta^{-\omega}$, $X_0 = \{0, 1, \ldots, m-1\}$, $c_0 \neq 0$.

For example, option (i) is used to describe the solution to equation (1) in the theory of discrete dynamic systems (see [6]). Option (ii) is most useful in problems of enumerative combinatorial analysis (see [17]).

In the case of constant coefficients, the z-transformation

$$F(z) = \sum_{x=0}^{\infty} \frac{f(x)}{z^x}$$

is the powerful method to study discrete dynamic systems and generating functions

$$F(z) = \sum_{x=0}^{\infty} f(x) z^x$$

are used for studying problems in enumerative combinatorial analysis.

In the multi-dimensional case, the number of formulations of Cauchy problem (1)–(2) increases. We discuss some of them.

An analogue of the one-dimensional case (ii), when $X = \mathbb{Z}^n_{\geqslant}$ is the non-negative octant in $\mathbb{Z}^n$, $X_0 = \mathbb{Z}^n_{\geqslant} \setminus X_m$, $0 \in \Omega$, $m_i = \max\{\omega_i : \omega_i \in \Omega,\ i = 1, \ldots, n\}$, $m = (m_1, \ldots, m_n)$ and $X_m = \{x \in \mathbb{Z}^n_{\geqslant} : x_i \geqslant m_i,\ i = 1, \ldots, m\}$, is considered in [2], which is devoted to multi-dimensional difference equations with constant coefficients and their use in enumerative combinatorial analysis. Several equivalent conditions, providing solvability of problem (1)–(2), are given in Theorem 3 in [2]. Particularly, the convex hull $\mathrm{conv}\{\Omega \setminus \{0\}\} \cap \mathbb{R}^n_{\geqslant}$ is not empty.

Various analogues of variant (i) of the Cauchy problem for the multi-dimensional case are constructed as follows. Let $A = \{\alpha^1, \ldots, \alpha^N\}$ be the set of vectors $\alpha^j = (\alpha^j_1, \ldots, \alpha^j_n) \in \mathbb{Z}^n$, $j = 1, \ldots, N$, and K is a lattice cone spanned by these vectors

$$K = \{x \in \mathbb{Z}^n : x = \lambda_1 \alpha^1 + \cdots + \lambda_N \alpha^N, \lambda_i \in \mathbb{Z}_{\geqslant}, i = 1, \ldots, N\}.$$

For points $u, v \in K$ a partial order relation $\underset{K}{\geqslant}$ is defined as follows: $u \underset{K}{\geqslant} v \Leftrightarrow u - v \in K$. We also denote $u \underset{K}{\not\geqslant} v \Leftrightarrow u - v \notin K$. We assume that the cone K is *pointed*, which means it does not contain any line or, equivalently, lies in an open half-space of $\mathbb{R}^n$.

We consider a finite set of integer points $A \subset K$, in which there exists a point m such that for all $\alpha^j \in A$, $j = 1, \ldots, N$ the condition $\alpha^j \underset{K}{\leqslant} m$ holds.

The solvability of the problem when the cone K is simplicial (which means that every element in it admits a unique expansion in the generators) and the sets $X = K$ and $X_0 = X \setminus (m + K)$, on which Cauchy problem (1)–(2) is solved, was studied in [1, 10–13, 19]. Additionally, in these papers, the solutions $f(x)$ to problem (1)–(2) are given in terms of the Cauchy data and fundamental solution to (1)–(2) (the Green function). These solutions play an important role in the study of asymptotics of solutions to the Cauchy problem, in particular, to study the stability of the problem and its connection with the properties of the characteristic set $\mathcal{V}_P := \{z \in \mathbb{C}^n : P(z) := \sum\limits_{\omega \in \Omega} c_\omega \, z^\omega \; = \; 0\}$ of the equation (1), where $z^\omega = z_1^{\omega_1} \cdot \ldots \cdot z_N^{\omega_N}$.

A multidimensional analogue of option (ii) for the Cauchy problem(1)–(2) was not described in [2]. This is apparently due to the fact that in problems of enumerative combinatorial analysis the search for the generating function for the combinatorial object is considered as a full solution to the problem, rather than the study of its asymptotic behavior.

For $n > 1$ we formulate the following variant of a Cauchy problem, which combines multidimensional analogs of (i) and (ii) for which the simplicity of the cone K is not required. We denote $m = \alpha^1 + \cdots + \alpha^N$, $c_0 = 1$, $\alpha^0 = (0, \ldots, 0)$.

The Cauchy problem. Find a function $f : K \to \mathbb{C}$, satisfying the difference equation

$$\sum_{j=0}^{N} c_j f(x - \alpha^j) = g(x), \quad x \underset{K}{\geqslant} m, \tag{3}$$

and which coincides with the given function $\varphi(x)$ on the set $X_0 = \{x \in K : x \underset{K}{\not\geqslant} m\}$:

$$f(x) = \varphi(x), \quad x \in X_0. \tag{4}$$

The characteristic polynomial for (3) is a Laurent polynomial (since it may have terms of negative degree) $P(z) = \sum\limits_{j=0}^{N} c_j z^{-\alpha^j}$.

Equation (3) with initial data (4) is used to describe a major class of problems in enumerative combinatorial analysis such as lattice path problems (the Dyck, Motzkin, Schröder and generalized lattice paths, see [2, 4, 15]).

The fact that the cone K is pointed allows us to use the method of generating functions. This involves defining for any $\mu \in K$ the element in the ring $\mathbb{C}_K[[z]]$ of (formal) power series

$$F_\mu(z) = \sum_{x \underset{K}{\geqslant} \mu} f(x) z^x.$$

We also define $F(z) = F_0(z)$.

Using the method of generating functions, we will derive a formula which expresses the generating function $F(z)$ in terms of the characteristic polynomial for (3) and generating functions for the Cauchy data.

Theorem 1. *The generating function $F(z)$ of a solution $f(x)$ to difference equation (3) with initial data (4) is representable as*

$$F(z) = \frac{1}{P(z^{-1})} \Big(\sum_{j=0}^{N} c_j z^{\alpha^j} \Phi_{m-\alpha^j}(z) + G_m(z) \Big), \tag{5}$$

where $P(z^{-1}) = P(z_1^{-1}, \ldots, z_n^{-1})$, $\Phi_{m-\alpha^j}(z) = F(z) - F_{m-\alpha^j}(z)$ and $G_m(z) = \sum\limits_{x \underset{K}{\geqslant} m} g(x) z^x$.

Proof. Multiplying the left-hand side of (3) by z^x and summing over $x \underset{K}{\geqslant} m$ yields

$$\sum_{x \underset{K}{\geqslant} m} \sum_{j=0}^{N} c_j f(x-\alpha^j) z^x = \sum_{j=0}^{N} c_j z^{\alpha^j} \sum_{x \underset{K}{\geqslant} m} f(x-\alpha^j) z^{x-\alpha^j} = \sum_{j=0}^{N} c_j z^{\alpha^j} \sum_{x+\alpha^j \underset{K}{\geqslant} m} f(x) z^x =$$

$$= \sum_{j=0}^{N} c_j z^{\alpha^j} F_{m-\alpha^j}(z) = \sum_{j=0}^{N} c_j z^{\alpha^j} \left(F(z) - \Phi_{m-\alpha^j}(z)\right).$$

Repeating the same with the right-hand side yields

$$P(z^{-1}) \cdot F(z) = \sum_{j=0}^{N} c_j z^{\alpha^j} \Phi_{m-\alpha^j}(z) + G_m(z).$$

Thus we obtain (5), which proves the theorem. □

Remark 1. Formulae (5) was derived in [14] for the Riordan arrays and in [11] for $K = \mathbb{Z}^N$ and $g(x) = 0$.

A function $\mathcal{P} : \mathbb{Z}^n \to \mathbb{C}$ is called *a fundamental solution* to the Cauchy problem (3)–(4) if it satisfies to the difference equation

$$\sum_{j=0}^{N} c_j \mathcal{P}(x-\alpha^j) = \delta_0(x), \quad x \in \mathbb{Z}^n, \tag{6}$$

where $\delta_0(x)$ is the Kronecker symbol:

$$\delta_0(x) = \begin{cases} 0 & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases}$$

The support of the function $\mathcal{P}(x)$ is a set

$$\operatorname{supp} \mathcal{P}(x) = \{x \in \mathbb{Z}^n : \mathcal{P}(x) \neq 0\}.$$

Lemma. If $\mathcal{P}(x)$ is the fundamental solution to Cauchy problem (3)–(4) and $\operatorname{supp} \mathcal{P} \subset K$, where K is a pointed cone, then

$$P(z^{-I}) \cdot \sum_{x \in \mathbb{Z}^n} \mathcal{P}(x) z^x = 1. \tag{7}$$

Proof. The product

$$\sum_{j=0}^{N} c_j z^{\alpha^j} \cdot \sum_{x \in \mathbb{Z}^n} \mathcal{P}(x) z^x = \sum_{j=0}^{N} \sum_{x \in \mathbb{Z}^n} c_j \mathcal{P}(x) z^{x+\alpha^j} = \sum_{x \in \mathbb{Z}^n} \sum_{j=0}^{N} c_j \mathcal{P}(x-\alpha^j) z^x = \sum_{x \in \mathbb{Z}^n} \delta_0(x) = 1,$$

which proves the lemma. □

The fundamental solution is

$$\mathcal{P}(x) = \sum_{\substack{A\lambda = x \\ \lambda \in \mathbb{Z}^N_{\geqslant}}} \frac{(-c_1)^{\lambda_1} \cdot \ldots \cdot (-c_N)^{\lambda_N} (\lambda_1 + \cdots + \lambda_N)!}{\lambda_1! \ldots \lambda_N!}, \quad x \underset{K}{\geqslant} 0,$$

and can be obtained by expanding $\dfrac{1}{P(z^{-1})}$ into the Laurent series as follows:

$$\frac{1}{P(z^{-1})} = \frac{1}{1 - \sum\limits_{j=1}^{N}(-c_j)z^{\alpha^j}} = \sum_{k=0}^{\infty}\left(\sum_{j=1}^{N}(-c_j)z^{\alpha^j}\right)^k =$$

$$= \sum_{\lambda_1+\ldots+\lambda_N \geqslant 0} \frac{(-c_1)^{\lambda_1}\cdot\ldots\cdot(-c_N)^{\lambda_N}(\lambda_1+\cdots+\lambda_N)!}{\lambda_1!\ldots\lambda_N!}\, z^{\lambda_1\alpha^1+\ldots+\lambda_N\alpha^N} =$$

$$= \sum_{x \underset{K}{\geqslant} 0} \sum_{\substack{A\lambda=x \\ \lambda\in\mathbb{Z}^N_{\geqslant}}} \frac{(-c_1)^{\lambda_1}\cdot\ldots\cdot(-c_N)^{\lambda_N}(\lambda_1+\cdots+\lambda_N)!}{\lambda_1!\ldots\lambda_N!}\, z^x = \sum_{x \underset{K}{\geqslant} 0} \mathcal{P}(x)\, z^x.$$

The Laurent series $\sum\limits_{x \underset{K}{\geqslant} 0} \mathcal{P}(x)\, z^x$ converges in a domain which can be described in term of an amoeba $\mathcal{A}_P$ of the Laurent polynomial $P(z)$. Namely, the logarithmic image of the domain is a complement component of the amoeba $\mathcal{A}_P$ corresponded with the point 0 of the Newtone polytope $\mathcal{N}_P$ (see [7]).

Function $P_A(x;h) = \sum\limits_{\substack{A\lambda=x \\ \lambda\in\mathbb{Z}^N_{\geqslant}}} h(\lambda)$ was considered in [15] and called *the vector partition function associated with* $h(\lambda)$. Provided that $h(\lambda) = \dfrac{(-c)^\lambda|\lambda|!}{\lambda!}$, we get

$$\mathcal{P}(x) = P_A(x;h). \tag{8}$$

For $h(\lambda) \equiv 1$ the vector partition function $P_A(x;h) = P_A(x)$ is a number of non-negative integer solutions to a linear Diophantine equation $A\lambda = x$ (see, for example, [17]):

$$P_A(x) = \sum_{\substack{A\lambda=x \\ \lambda\in\mathbb{Z}^N_{\geqslant}}} 1, \quad x \in \mathbb{Z}^n.$$

For $h(\lambda) = e^{-\langle\lambda,y\rangle}$ properties of the function

$$P_A(y;h) = \sum_{\substack{A\lambda=x \\ \lambda\in\mathbb{Z}^N_{\geqslant}}} e^{-\langle\lambda,y\rangle}, \; y \in \mathbb{C}^N, \tag{9}$$

called *the vector partition function associated with the set of vectors* A, were investigated in [3]. In particular, they derive the residue formulas for its generating function and an analog of the Euler-Maclaurin formula, in which the vector partition functions are represented as the action of the Todd operator on the volume function of a polyhedron. Furthermore, a sum of $e^{-\langle\lambda,y\rangle}$ in integer cones was investigated in [16] in connection to generalization of the Riemann-Roch theorem. A structure theorem for the vector partition function was presented and polyhedral tools for the efficient computation of such functions was provided in [18].

For $\varphi(x) \equiv 1$ the function $P_A(\lambda;\varphi)$ coincides with the classical vector partition function. For $\varphi(x) = e^{-\langle x,y\rangle}$ we obtain a vector partition function of the form (9). If we take $N = 2$, $A = (1\ 1)$ and $\varphi(x_1,x_2) = h(x_1)$, then $P_A(\lambda;\varphi) = \sum\limits_{\substack{x_1+x_2=\lambda \\ x_1,x_2\geqslant 0}} h(x_1) = \sum\limits_{x_1=0}^{\lambda} h(x_1)$. Thus, the problem of finding the vector partition function $P_A(\lambda;\varphi)$ is a generalization of the classical summation's problem of functions of a discrete argument.

The concept of the fundamental solution $\mathcal{P}(x)$ to (3)–(4) yields a formula expressing $f(x)$ in terms of Cauchy data $\varphi(x)$ and $g(x)$.

Theorem 2. *A solution to the difference equation (3) with initial data (4) is given as follows:*

$$f(x) = \sum_{0 \underset{K}{\leqslant} y \underset{K}{\leqslant} x} \mathcal{P}(x-y)\tau(y),$$

$$\text{where } \tau(y) = \begin{cases} \sum\limits_{j=0}^{N} c_j \varphi(y-\alpha^j) & \text{if } y \underset{K}{\not\geqslant} m, \\ g(y) & \text{if } y \underset{K}{\geqslant} m. \end{cases}$$

Proof. Using expression (5) from Theorem 1 and expression (7) from Lemma yields

$$F(z) = \sum_{x \underset{K}{\geqslant} 0} \mathcal{P}(x) z^x \left(\sum_{j=0}^{N} c_j z^{\alpha^j} \Phi_{m-\alpha^j}(z) + G_m(z) \right).$$

Since

$$\sum_{j=0}^{N} c_j z^{\alpha^j} \Phi_{m-\alpha^j}(z) = \sum_{y \underset{K}{\geqslant} 0, y \underset{K}{\not\geqslant} m} \left(\sum_{j=0}^{N} c_j \varphi(y-\alpha^j) \right) z^y,$$

we get

$$F(z) = \sum_{x \underset{K}{\geqslant} 0} \mathcal{P}(x) z^x \sum_{y \underset{K}{\geqslant} 0} \tau(y) z^y,$$

$$\text{where } \tau(y) = \begin{cases} \sum\limits_{j=0}^{N} c_j \varphi(y-\alpha^j) & \text{if } y \underset{K}{\not\geqslant} m, \\ g(y) & \text{if } y \underset{K}{\geqslant} m. \end{cases}$$

Finally, taking into account that $\mathcal{P}(x) = 0$ for $x \underset{K}{\not\geqslant} 0$ we get

$$F(z) = \sum_{x \underset{K}{\geqslant} 0} \left(\sum_{y \underset{K}{\geqslant} 0} \mathcal{P}(x)\tau(y) \right) z^{x+y} = \sum_{x \underset{K}{\geqslant} 0} \left(\sum_{0 \underset{K}{\leqslant} y \underset{K}{\leqslant} x} \mathcal{P}(x-y)\tau(y) \right) z^x.$$

Equating the coefficients of z^x we obtain

$$f(x) = \sum_{0 \underset{K}{\leqslant} y \underset{K}{\leqslant} x} \mathcal{P}(x-y)\tau(y),$$

which proves the theorem. □

2. Applications to lattice path problems

A lattice path is a finite sequence $p_0, p_1, \ldots, p_L$ of points in $\mathbb{Z}^n$ and its steps are the finite set of lattice vectors $p_k - p_{k-1} \in A = \{\alpha^1, \ldots, \alpha^N\}$, $k = 1, 2, \ldots, L$. The common class of lattice paths arises by imposing some conditions on the paths: points p_k, $k = 0, 1, \ldots, L$, are distinct (non intersecting paths). In the context of lattice path counting problems the function $f : \mathbb{Z}^N \to \mathbb{Z}_{\geqslant}$ that counts the number $f(x)$ of paths in a specified class for which $p_0 = 0$ is computed (the

condition $p_0 = 0$ does not result in a loss of generality). Examples of some well-known lattice paths: Dyck, Motzkin and Schröder paths (for more details see [2, 5, 8, 9]).

It is well-known that the function $f(x)$ satisfies difference equation (3) with $c_0 = 1$, $c_1 = \ldots = c_N = -1$ and $g(x) = 0$ (see [2]). Thus $P(\delta) = 1 - \delta^{-\alpha^1} - \cdots - \delta^{-\alpha^N}$.

Theorem 2 yields a simple formula for the number $f(x)$ of such paths (see also [15]). The following condition for an initial data function $\varphi(x)$ of Cauchy problem (3)–(4) for the lattice path problem holds:

$$\varphi(x) = \begin{cases} 0 & \text{if } x \not\geqslant\limits_K 0, \\ 1 & \text{if } x = 0, \\ (1 - P(\delta))\varphi(x) & \text{if } x \geqslant\limits_K 0, x \neq 0. \end{cases}$$

Since $\tau(y)$ is equal to 1 only at the origin and vanishes at other points we get $f(x) = \mathcal{P}(x)$. Considering (8) we obtain

$$f(x) = P_A(x; h), \text{ where } h(\lambda) = \frac{|\lambda|!}{\lambda!}.$$

Example A.

We consider a set with three steps $A = \{\alpha^1 = (1, 0), \alpha^2 = (0, 1), \alpha^3 = (1, 1)\}$ and let $f(x_1, x_2)$ denote the number of paths from the origin to $(x_1, x_2) \in \mathbb{Z}^2$ using steps from the set A. The cone K is spanned by the vectors from A and $m = \alpha_1 + \alpha_2$, since $\alpha_3 = \alpha_1 + \alpha_2$.

We consider the two dimensional difference equation

$$f(x_1, x_2) - f(x_1 - 1, x_2) - f(x_1, x_2 - 1) - f(x_1 - 1, x_2 - 1) = 0, \tag{10}$$

and its characteristic polynomial $P(z_1, z_2) = 1 - z_1^{-1} - z_2^{-1} - z_1^{-1} z_2^{-1}$.

By Theorem 2 a solution to this difference equation is

$$f(x_1, x_2) = \sum_{0 \leqslant\limits_K y \leqslant\limits_K x} \mathcal{P}(x_1 - y_1, x_2 - y_2)\tau(y_1, y_2),$$

where $\tau(y_1, y_2) = \varphi(y_1, y_2) - \varphi(y_1 - 1, y_2) - \varphi(y_1, y_2 - 1) - \varphi(y_1 - 1, y_2 - 1)$ if $(y_1, y_2) \not\geqslant (1, 1)$ and $\tau(y_1, y_2) = 0$ otherwise.

To find the fundamental solution $\mathcal{P}(x_1, x_2)$ we expand $P^{-1}(z_1^{-1}, z_2^{-1})$ as follows

$$\frac{1}{P(z_1^{-1}, z_2^{-1})} = \frac{1}{1 - (z_1 + z_2 + z_1 z_2)} = \sum_{k=0}^{\infty} (z_1 + z_2 + z_1 z_2)^k =$$

$$= \sum_{k_1, k_2, k_3 \geqslant 0} \frac{(k_1 + k_2 + k_3)!}{k_1! k_2! k_3!} z_1^{k_1} z_2^{k_2} (z_1 z_2)^{k_3} = \sum_{x_1, x_2 \geqslant 0} \sum_{t=0}^{min(x_1, x_2)} \frac{(x_1 + x_2 - t)!}{(x_1 - t)!(x_2 - t)! t!} z_1^{x_1} z_2^{x_2}.$$

Consequently, Lemma and the term of the fundamental solution gives

$$\mathcal{P}(x_1, x_2) = \sum_{t=0}^{min(x_1, x_2)} \frac{(x_1 + x_2 - t)!}{(x_1 - t)!(x_2 - t)! t!} = \sum_{\substack{k_1 + k_3 = x_1 \\ k_2 + k_3 = x_2 \\ k_1, k_2, k_3 \geqslant 0}} \frac{(k_1 + k_2 + k_3)!}{k_1! k_2! k_3!} = P_A(x; \lambda).$$

Finally, we have the solution for difference equation (10) with initial data function $f(x_1, x_2) = \varphi(x_1, x_2)$, $(x_1, x_2) \not\geqslant (1, 1)$ as follows

$$f(x_1, x_2) = \mathcal{P}(x_1, x_2)\varphi(0, 0) + \sum_{y_1=1}^{x_1} \mathcal{P}(x_1 - y_1, x_2)(\varphi(y_1, 0) - \varphi(y_1 - 1, 0)) + \\ + \sum_{y_2=1}^{x_2} \mathcal{P}(x_1, x_2 - y_2)(\varphi(0, y_2) - \varphi(0, y_2 - 1)).$$

In the case of lattice paths, $\varphi(y_1, 0) - \varphi(y_1 - 1, 0) = 0$ for $y_1 \geqslant 1$, $\varphi(0, y_2) - \varphi(0, y_2 - 1) = 0$ for $y_2 \geqslant 1$, and $\varphi(0, 0) = 1$, we obtain

$$f(x_1, x_2) = \mathcal{P}(x_1, x_2).$$

Example B.

Let $\alpha^1 = (2, -1)$, $\alpha^2 = (-1, 2)$ be a column vectors, we let K denote the cone K spanned by the vectors $K = \langle \alpha^1, \alpha^2 \rangle$, $m = \alpha^1 + \alpha^2 = (1, 1)$.

We consider the two dimensional difference equation

$$f(x_1, x_2) - f(x_1 - 2, x_2 + 1) - f(x_1 + 1, x_2 - 2) = 0 \tag{11}$$

and its characteristic polynomial $P(z_1, z_2) = 1 - z_1^{-2} z_2 - z_1 z_2^{-2}$.

By Theorem 2 a solution to this difference equation is

$$f(x_1, x_2) = \sum_{0 \underset{K}{\leqslant} y \underset{K}{\leqslant} x} \mathcal{P}(x_1 - y_1, x_2 - y_2)\tau(y_1, y_2),$$

where $\tau(y_1, y_2) = \begin{cases} \varphi(y_1, y_2) - \varphi(y_1 - 2, y_2 + 1) - \varphi(y_1 + 1, y_2 - 2) & \text{if } (y_1, y_2) \underset{K}{\not\geqslant} (1, 1), \\ 0 \text{ if } (y_1, y_2) \underset{K}{\geqslant} (1, 1). \end{cases}$

To find a fundamental solution $\mathcal{P}(x_1, x_2)$ we expand the characteristic polynomial $P(z_1^{-1}, z_2^{-1})$ into a series:

$$\frac{1}{1 - z_1^2 z_2^{-1} - z_1^{-1} z_2^2} = \sum_{k=0}^{\infty} (z_1^2 z_2^{-1} + z_1^{-1} z_2^2)^k = \sum_{k_1 + k_2 \geqslant 0} \frac{(k_1 + k_2)!}{k_1! k_2!} (z_1^2 z_2^{-1})^{k_1} (z_1^{-1} z_2^2)^{k_2} = \\ = \sum_{k_1 + k_2 \geqslant 0} \frac{(k_1 + k_2)!}{k_1! k_2!} z_1^{2k_1 - k_2} z_2^{-k_1 + 2k_2} = \sum_{(x_1, x_2) \underset{K}{\geqslant} 0} \frac{(x_1 + x_2)!}{\left(\frac{2x_1 + x_2}{3}\right)! \left(\frac{x_1 + 2x_2}{3}\right)!} z_1^{x_1} z_2^{x_2}.$$

Consequently,

$$\mathcal{P}(x_1, x_2) = \frac{(x_1 + x_2)!}{\left(\frac{2x_1 + x_2}{3}\right)! \left(\frac{x_1 + 2x_2}{3}\right)!}.$$

Finally, we have the solution for difference equation (11) with arbitrary initial data

$$f(x_1, x_2) = \varphi(x_1, x_2), \quad (x_1, x_2) \underset{K}{\not\geqslant} (1, 1)$$

$$f(x_1, x_2) = \mathcal{P}(x_1, x_2)\varphi(0, 0) + \sum_{t=1}^{x_1} \mathcal{P}(x_1 - 2t, x_2 + t)(\varphi(2t, -t) - \varphi(2t - 2, -t + 1)) + \\ + \sum_{t=1}^{x_2} \mathcal{P}(x_1 + t, x_2 - 2t)(\varphi(-t, 2t) - \varphi(-t + 1, 2t - 2)).$$

In the case of lattice paths, $\varphi(2t,-t)-\varphi(2t-2,-t+1)=0$ for $t \geqslant 1$, $\varphi(-t,2t)-\varphi(-t+1,2t-2)=$ $=0$ for $t \geqslant 1$, and $\varphi(0,0)=1$, we obtain

$$f(x_1,x_2)=\mathcal{P}(x_1,x_2).$$

This work of author was financed by the PhD SibFU grant for support of scientific research no. 14.

References

[1] M.S.Apanovich, E.K.Leinartas, *J. Sib. Fed. Univ. Math. Phys*, **10**(2017), no. 2, 199–205. DOI: 10.26516/1997-7670.2018.26.3

[2] M.Bousquet-Mélou, M.Petkovšek, *Discrete Mathematics*, **225**(2000), 51–75.

[3] M.Brion, M.Vergne, *J. American Math. Soc.*, Vol. 10, no. 4 (1997), pp. 797–833. DOI: 10.1090/S0894-0347-97-00242-7

[4] S.Chandragiri, *J. Sib. Fed. Univ. Math. Phys*, **12**(2019), no. 5, 551–559. DOI: 10.17516/1997-1397-2019-12-5-551-559

[5] P.Duchon, *Discrete Math*, **225**(2000), 121–135.

[6] S.N.Elaydi, An Introduction to Difference Equations, 3rd edn. Undergraduate Texts in Mathematics, Springer, New York, 2005.

[7] M.Forsberg, M.Passare, A.Tsikh, *Advances in Mathematics*, 151(2000), no. 1, 45–70.

[8] I.M.Gessel, *J. Combin. Theory. Ser. A*, 28(1980), 321–337.

[9] J.Labelle, Y.N.Yeh, Generalized Dyck paths, *Discrete Math*, 82(1990), 1–6.

[10] E.K.Leinartas, *Siberian Mathematical Journal*, **48**(2007), no. 2, 268–272. DOI: 10.1007/s11202-007-0026-0

[11] E.K.Leinartas, A.P.Lyapin, *J. Sib. Fed. Univ. Math. Phys*, **2**(2009), no. 4, 449–455 (in Russian).

[12] E.K.Leinartas, T.I.Nekrasova, *Siberian Mathematical Journal*, **57**(2016), no. 2, 98–112. DOI: 10.17377/smzh.2016.57.108

[13] E.K.Leinartas, M.S.Rogozina, *Siberian Mathematical Journal*, **56**(2015), no. 1, 92–100. DOI: 10.1134/S0037446615010097

[14] A.P.Lyapin, *J. Sib. Fed. Univ. Math. Phys*, **2**(2009), no. 2, 210–220.

[15] A.P.Lyapin, S.Chandragiri, *Journal of Difference Equations and Applications*, **25**(2019), no. 7, 1052–1061. DOI: 10.1080/10236198.2019.1649396

[16] A.V.Pukhlikov, A.G.Khovanskii, *St. Petersburg Mathematical Journal*, **4**(1993), no. 4, 789–812.

[17] R.Stanley, Enumerative combinatorics, Cambridge Univ. Press, Cambridge, 1999.

[18] B.Sturmfels, *Journal of Combinatorial Theory. Series A*, **72**(1995), 302–309.
DOI: 10.1016/0097-3165(95)90067-5

[19] T.I.Yakovleva, *Siberian Mathematical Journal*, **58**(2017), no. 2, 363–372.
DOI: 10.1134/S0037446617020185

Задача Коши для многомерного разностного уравнения в конусах целочисленной решетки

Александр П. Ляпин
Сибирский федеральный университет
Красноярск, Российская Федерация
Лесосибирский педагогический институт — филиал СФУ
Лесосибирск, Российская Федерация
Шрилатха Чандрагири
Сибирский федеральный университет
Красноярск, Российская Федерация

Аннотация. В работе рассмотрен вариант задачи Коши для многомерного разностного уравнения с постоянными коэффициентами, возникающий с задачей о числе путей на целочисленной решетке в перечислительном комбинаторном анализе. Получена формула, выражающая производящую функцию решения задачи Коши через производящие функции данных Коши, и найдено решение задачи Коши через ее фундаментальное решение и данные Коши.

Ключевые слова: разностное уравнение, фундаментальное решение, производящая функция, пути Дика.

DOI: 10.17516/1997-1397-2020-13-2-197-212
УДК 519.21

Rotationally-axisymmetric Motion of a Binary Mixture with a Flat Free Boundary at Small Marangoni Numbers

Victor K. Andreev*
Institute of Computational Modelling SB RAS
Krasnoyarsk, Russian Federation
Siberian Federal University
Krasnoyarsk, Russian Federation
Natalya L. Sobachkina†
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 06.092019, received in revised form 06.11. 2019, accepted 06.02.2020

Abstract. Rotationally-axisymmetric motion of a binary mixture with a flat free boundary at small Marangoni numbers is investigated. The problem is reduced to the inverse linear initial-boundary value problem for parabolic equations. Using Laplace transformation properties the exact analytical solution is obtained. It is shown that a stationary solution is the limiting one with the growth of time if there is a certain relationship between the temperature of the solid wall and the external temperature of the gas. If there is no connection, the convergence to the stationary solution is broken. Some examples of numerical reconstruction of the temperature, concentration and velocity fields are given, which confirm the theoretical conclusions.

Keywords: binary mixture, free boundary, inverse problem, the pressure gradient, the stationary solution, Laplace transformation, thermal Marangoni number.

Citation: V.K.Andreev, N.L.Sobachkina, Rotationally-axisymmetric Motion of a Binary Mixture with a Flat Free Boundary at Small Marangoni Numbers, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 197-212. DOI: 10.17516/1997-1397-2020-13-2-197-212.

Introduction

The main purpose of this work is to construct an exact solution of the inverse initial boundary value problem of rotationally symmetric motion of a viscous heat-conducting binary mixture with a flat free boundary at small Marangoni numbers, as well as a numerical solution of the problem.The movement is caused by a non-stationary pressure gradient.

It is well known that for small Marangoni numbers, the momentum equation can be simplified by discarding convective acceleration. Such movements are called *crawling*. Similar simplifications can be obtained for the energy and concentration transfer equations. One of these problems, considered in paper [1], is devoted to the study of solving the thermodiffusion equations of a special type that describes the two-dimensional motion of a binary mixture in a flat channel. In the resulting initial boundary value problem, the analog of the Marangoni number is the Reynolds number. Assuming that this number is small, the problem becomes linear. Its solution is found using trigonometric Fourier series that converge rapidly for any given time.

There are a lot of theoretical works concerning convective movements in flat layers with a free boundary. R. V. Birikh's *exact stationary solutions* to the problem of thermocapillary convection

*andr@icm.krasn.ru
†sobachkinanat@mail.ru

in a flat horizontal layer are well known in work [2]. One solution describes the flow in the band $-h < x < 0$, both borders of which are solid walls, and in the second — the upper border of the band is free, subject to the action of thermocapillary forces. The solutions were widely used and cited [3–15]. In a number of these works [6,7,10,13–15], the flat Benard-Marangoni convection of a viscous incompressible liquid was studied in the Oberbeck–Bussinesque model. A characteristic feature of the obtained solutions is the one-dimensional velocity coordinates, and the temperature and pressure fields are three-dimensional. In the work [13], an exact solution was obtained near the point of the temperature extremum at zero Grasgoff number. The found solution serves as an initial approximation for constructing solutions for Grasshoff numbers greater than zero. In works [14,15] of the initial boundary value problem describing non-stationary layered flows of the Benard–Marangoni convection in an infinitely extended flat layer, the existence of counterflows in the liquid layer was found. The presence of counterflows is equivalent to the presence of stagnant points, which indicates the existence of a local extremum of the kinetic energy of the liquid.

In this paper, in the absence of external forces, we study the creeping axisymmetric motion of a mixture with a flat free boundary with a Hiemenz type velocity field [16]. Here *the inverse problem* arises, since the non-stationary pressure gradient is also the desired function.

1. Statement of the problem

We consider the axisymmetric motion of an infinite horizontal plane layer of a viscous heat-conducting binary mixture bounded by a solid wall $z = 0$ and a free boundary $z = l(t)$ (see Fig. 1). Let $\mathbf{u}(\mathbf{x}, t)$ is the velocity vector, $p(\mathbf{x}, t)$ is the pressure, $\theta(\mathbf{x}, t)$, $c(\mathbf{x}, t)$ are deviations from the average values temperature and concentration values of the mixture under conditions of complete weightlessness. The process is described by a system of equations of thermodiffusion motion [17]:

$$\begin{aligned} &\frac{d\mathbf{u}}{dt} + \frac{1}{\rho}\nabla p = \nu\Delta\mathbf{u}, \quad \operatorname{div}\mathbf{u} = 0, \\ &\frac{d\theta}{dt} = \chi\Delta\theta, \quad \frac{dc}{dt} = d\Delta c + \alpha d\Delta\theta, \end{aligned} \tag{1}$$

where ρ is the average density, ν is the kinematic viscosity, χ is the thermal diffusivity, d is the diffusion coefficient, α is the thermodiffusion coefficient (Soret coefficient); $d/dt = \partial/\partial t + \mathbf{u}\cdot\nabla$ is the full time derivative, Δ is the Laplace operator.

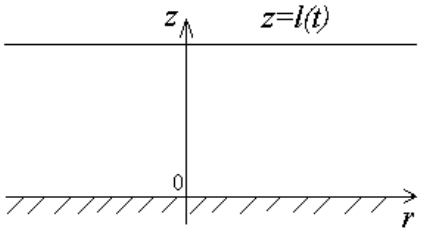

Fig. 1. Diagram of the flow region

Remark 1. The equation of energy from the system (1) does not take into account the term describing the dissipation of kinetic energy. This is due to the fact that the ratio of this term and $\mathbf{u}\cdot\nabla\theta$ for most processes does not exceed 10^{-7}. In addition, all model parameters are assumed to be constant, and they are reliably determined experimentally.

Let $u(r, z, t)$, $w(r, z, t)$ are projections of the velocity vector in cylindrical coordinate system.

The solution of the problem is searched for in a special form:

$$u = ru_1(z,t), \quad w = w(z,t), \quad p = p(r,z,t), \quad \theta = a(z,t)r^2 + b(z,t), \\ c = h(z,t)r^2 + g(z,t). \tag{2}$$

A solution of the form (2) is called a Hiemenz type solution [16], in which the velocity field is linear relative to one from the coordinates. It is partially invariant with respect to the five-parameter subgroup generated by the operators $\partial/\partial r,\ t\partial/\partial r + \partial/\partial u,\ \partial/\partial\theta,\ \partial/\partial c,\ \partial/\partial p$ [18].

Substituting the form (2) into the system of thermodiffusion equations leads to the system (reassign $u_1 \leftrightarrow u$):

$$ru_t + ru^2 + rwu_z + \frac{1}{\rho}\,p_r = r\nu u_{zz}; \tag{3}$$

$$w_t + ww_z + \frac{1}{\rho}\,p_z = \nu w_{zz}; \tag{4}$$

$$2u + w_z = 0; \tag{5}$$

$$a_t + 2au + wa_z = \chi a_{zz}; \tag{6}$$

$$b_t + wb_z = \chi(4a + b_{zz}); \tag{7}$$

$$h_t + 2hu + wh_z = dh_{zz} + \alpha d a_{zz}; \tag{8}$$

$$g_t + wg_z = d(4h + g_{zz}) + \alpha d(4a + b_{zz}), \tag{9}$$

that needs to be solved in the field $t > 0,\ 0 < z < l(t)$.

It is assumed that the surface tension coefficient σ at the free boundary linearly depends on the temperature and concentration

$$\sigma(\theta, c) = \sigma_0 - æ_1(\theta - \theta_0) - æ_2(c - c_0),$$

where $æ_1 > 0$ is the temperature coefficient of surface tension, $æ_2$ is the concentration coefficient of surface tension (usually $æ_2 < 0$, since the surface tension increases with increasing concentration); θ_0, c_0 are some constant average values.

Boundary conditions on an unknown free boundary $z = l(t)$ for the system (3)–(9) have the form:

$$\frac{dl}{dt} = w(l(t), t); \tag{10}$$

$$u_z = -\frac{2æ_1}{\rho\nu}\,a - \frac{2æ_2}{\rho\nu}\,h; \tag{11}$$

$$p_{gas} - p + 2\rho\nu w_z = 0; \tag{12}$$

$$ka_z + \gamma(a - a_{gas}) = 0; \tag{13}$$

$$kb_z + \gamma(b - b_{gas}) = 0; \tag{14}$$

$$h_z + \alpha a_z = 0; \tag{15}$$

$$g_z + \alpha b_z = 0, \tag{16}$$

where p_{gas}, θ_{gas} are the pressure and the temperature of the surrounding gas; k, γ are the thermal conductivity and the heat transfer coefficients. It is assumed that the transfer processes in gas can be neglected. It is assumed that the gas pressure p_{gas} is constant, and its temperature θ_{gas} at the border with the liquid mixture is set by the function of time. Thus, the ratio (10) is the kinematic condition, (11), (12) are tangential and normal dynamic conditions, (13), (14) is a condition for heat exchange with the gas surrounding the mixture, (15), (16) is a condition for the absence of a flow of matter across a free boundary (thus the effect of surfactants on $z = l(t)$ is not taken into account).

Boundary conditions on a solid wall $z = 0$:

$$\begin{gathered} u(0,t) = 0\,, \quad w(0,t) = 0\,, \quad a(0,t) = a(t)\,, \quad b(0,t) = b(t), \\ h_z(0,t) + \alpha a_z(0,t) = 0\,, \quad g_z(0,t) + \alpha b_z(0,t) = 0. \end{gathered} \tag{17}$$

Initial conditions:

$$\begin{gathered} u(z,0) = u_0(z), \quad w(z,0) = w_0(z), \quad a(z,0) = a_0(z), 1 \quad b(z,0) = b_0(z), \\ h(z,0) = h_0(z), \quad g(z,0) = g_0(z), \quad l(0) = l_0 > 0, \end{gathered} \tag{18}$$

and the functions u_0, w_0, a_0, b_0 satisfy the conditions (17); u_0 and w_0 are connected by equation (5); $u_0\,, a_0$, h_0 are connected by condition (11); h_0, a_0 — by conditions (15) and (17); g_0, b_0 — by conditions (16) and (17). Thus, the approval conditions are met.

From the equations (3), (4), the pressure gradient (p_r, p_z) is expressed:

$$p_r = -r\rho(u_t + u^2 + wu_z - \nu u_{zz}); \tag{19}$$

$$p_z = \rho(w_{zz} - w_t - ww_z); \tag{20}$$

The compatibility conditions of the equations (19), (20) are satisfied identically: $p_{rz} = p_{zr} = 0$. It follows that the function $u(z,t)$ will be determined from the equation

$$u_t + u^2 + wu_z = \nu u_{zz} + f(t), \tag{21}$$

and the pressure is restored by the formula

$$p = -\frac{r^2}{2}\rho f(t) + s(z,t), \tag{22}$$

here $f(t)$ is arbitrary function, and the derivative of the variable z from the function $s(z,t)$ is exactly the right side of the equation (20). The function $s(z,t)$ is considered known if the function $w(z,t)$ is found.

Therefore, the problem is inverse, since the longitudinal pressure gradient $f(t)$ is an unknown function. In the theory of inverse problems, it is called a source function.

2. Converting to a task in a fixed area

You can see that the equations (21), (5), (6), (8) are independent of the others. They form a closed initial boundary value problem for defining the functions $u(z,t)$, $a(z,t)$, $h(z,t)$, and $l(t)$. Therefore, we will reduce the task to finding only these functions. To do this, we integrate the equation (5) and exclude the function w in the equations (21), (6), (8). In the resulting system, we introduce dimensionless variables and functions with equalities:

$$\begin{gathered} \tau = \frac{\nu t}{l_0^2}\,, \quad y = \frac{z}{l(t)}\,, \quad U = \frac{l_0^2 u}{\nu}\,, \quad A = \frac{l_0^2 a}{\bar{T}}, \\ A_{gas} = \frac{l_0^2 a_{gas}}{\bar{T}}\,, \quad H(z,t) = \frac{l_0^2 h}{\bar{c}}\,, \quad L(\tau) = \frac{l(t)}{l_0}\,, \quad F(\tau) = \frac{l_0^4 f(t)}{\nu^2}\,, \end{gathered} \tag{23}$$

here $\bar{T}$, $\bar{c}$ are characteristic temperature and concentration.

The result is a task in a fixed area $0 < y < 1$:

$$M(U) \equiv U_\tau - (\ln L)_\tau y U_y - 2U_y \int_0^y U(y,\tau)\,dy + U^2 - \frac{1}{L^2}\,U_{yy} - F(\tau) = 0; \tag{24}$$

$$F(U,A) \equiv A_\tau - (\ln L)_\tau y A_y - 2A_y \int_0^y U(y,\tau)\,dy + 2AU - \frac{1}{\mathrm{Pr}\,L^2}\,A_{yy} = 0; \tag{25}$$

$$\begin{gathered} R(U,A,H) \equiv H_\tau - (\ln L)_\tau y H_y - 2H_y \int_0^y U(y,\tau)\,dy + 2HU - \\ - \frac{1}{\mathrm{Sc}\,L^2}\,H_{yy} - \frac{\mathrm{Sr}}{\mathrm{Sc}\,L^2}\,A_{yy} = 0. \end{gathered} \tag{26}$$

In (24)–(26), dimensionless parameters are entered: $\mathrm{Sc} = \nu/d$ is Schmidt number, $\mathrm{Sr} = \alpha d\bar{T}/\nu\bar{c}$ is Soret number, $\mathrm{Pr} = \nu/\chi$ is Prandtl number.

The following conditions are met on a solid wall $y = 0$:

$$U(0,\tau) = 0, \quad A(0,\tau) = A(\tau), \quad H_y(0,\tau) + \mathrm{Sr}A_y(0,\tau) = 0. \tag{27}$$

On a free boundary $y = 1$:

$$\frac{dL}{d\tau} = -2L \int_0^1 U(y,\tau)\,dy; \tag{28}$$

$$A_y + L\,\mathrm{Bi}(A - A_{gas}) = 0; \tag{29}$$

$$-\frac{1}{2L}\,U_y = \mathrm{Ma}\,A + \mathrm{Mc}\,H; \tag{30}$$

$$H_y + \mathrm{Sr}A_y = 0, \tag{31}$$

where $\mathrm{Bi} = \gamma l_0/k$ is the number of Bio; $\mathrm{Ma} = æ_1\bar{T}l_0/\rho\nu^2$, $\mathrm{Mc} = æ_2\bar{c}l_0/\rho\nu^2$, respectively, the thermal Marangoni number and the concentration Marangoni number.

Initial conditions for $\tau = 0$:

$$\begin{gathered} U(y,0) = U_0(y)\,, \quad A(y,0) = A_0(y)\,, \quad H(y,0) = H_0(y)\,, \\ L(0) = 1\,, \quad F(0) = F^0 \equiv \mathrm{const}. \end{gathered} \tag{32}$$

To find an unknown pressure gradient $F(\tau)$ when solving the inverse problem, you need to set an additional condition. As such the condition is an integral redefinition condition, which is written as:

$$\int_0^1 U dy = 0, \quad y = 1. \tag{33}$$

This is a condition of closed flow. Thus, the flow rate of the liquid mixture through any normal cross-section is zero. Given the conditions (28) and (32), it follows from the integral redefinition condition (33) that the free boundary remains fixed and is equal to $L(\tau) = 1$.

3. Stationary solution

We will assume that the thermal Marangoni number is $\mathrm{Ma} \ll 1$ (*the creeping motion*), as well as $\mathrm{Ma} \sim \mathrm{Mc}$, that is, thermal and concentration effects on a free boundary of the same order. Formally decomposing the functions U, A, H in a row by Ma, we get for the first approximation the problem (24)–(26) with $\mathrm{Ma} = 0$. In the equations of momentum, heat transfer, and concentration, the convective terms are discarded. We will consider the steady flow of the liquid. For such a movement, all the required functions do not depend on time; let's denote them by $U^0(y)$, $A^0(y)$, $H^0(y)$, F^0. Also, on a solid wall, $A(\tau) = A \equiv \mathrm{const}$. Let's write out the corresponding boundary value problem for $0 < y < 1$, which becomes linear for small Marangoni numbers:

$$U^0_{yy} + F^0 = 0; \tag{34}$$

$$A^0_{yy} = 0; \tag{35}$$

$$H^0_{yy} + \mathrm{Sr} A^0_{yy} = 0, \tag{36}$$

with boundary conditions (27)–(31).

When searching for a stationary solution, a fundamental result was obtained. That is, in order for the solutions found to satisfy all boundary conditions, it is necessary and sufficient that the temperature of the solid wall is associated with the external temperature of the gas by a certain condition. The relationship between temperatures is as follows:

$$A = -\frac{\mathrm{Bi}\, A^0_{gas}}{\mathrm{Bi}+2}. \tag{37}$$

Then the required functions in the first approximation have the form:

$$A^0(y) = \frac{\mathrm{Bi}\, A^0_{gas}\,(2y-1)}{\mathrm{Bi}+2}; \tag{38}$$

$$H^0(y) = \frac{\mathrm{Bi}\,\mathrm{Sr}\, A^0_{gas}\,(1-2y)}{\mathrm{Bi}+2}; \tag{39}$$

$$U^0(y) = \frac{\mathrm{Bi}\, A^0_{gas}\,(1-\mathrm{MSr})(y-1,5y^2)}{\mathrm{Bi}+2}; \tag{40}$$

$$F^0 = 3\,\frac{\mathrm{Bi}\, A^0_{gas}\,(1-\mathrm{MSr})}{\mathrm{Bi}+2}, \tag{41}$$

where $\mathrm{M} = æ_2\bar{c}/æ_1\bar{T}$ is a dimensionless parameter equal to the ratio of the thermal Marangoni number to the concentration Marangoni number.

In addition, representations are found for other functions of the General problem, which made a significant contribution to obtaining a certain relationship between temperatures:

$$B^0(y) = -\frac{2}{3}\alpha_1 y^3 - 2\alpha_2 y^2 + \beta_1 y + \beta_2; \tag{42}$$

$$G^0(y) = -\frac{2}{3}\gamma_1 y^3 - 2\gamma_2 y^2 + \delta_1 y + \delta_2; \tag{43}$$

where α_1, α_2, β_1, β_2, γ_1, γ_2, δ_1, δ_2 are constants defined from boundary conditions (27)–(32):

$$\begin{gathered}\alpha_2 = -\frac{\mathrm{Bi}\, A^0_{gas}}{\mathrm{Bi}+2}, \quad \alpha_1 = -2\alpha_2, \quad \beta_2 = B, \\ \beta_1 = \frac{\mathrm{Bi}(B^0_{gas} - \beta_2 + \frac{2}{3}\alpha_1 + 2\alpha_2) + 2\alpha_1 + 4\alpha_2}{\mathrm{Bi}+1}, \\ \gamma_1 = -\alpha_1\,\mathrm{Sr}, \quad \gamma_2 = -\frac{\gamma_1}{2}, \quad \delta_1 = -\beta_1\,\mathrm{Sr}, \quad \delta_2 = \frac{\alpha_1\,\mathrm{Sr}}{6} + \frac{\beta_1\,\mathrm{Sr}}{2} + C.\end{gathered} \tag{44}$$

Here B is the second component of the solid wall temperature for the stationary case, and C is a constant that sets the average cross-section concentration $y = 0$.

4. Determining of the temperature field

For solution of nonstationary linear problem is used Laplace transform. Believe (assuming the existence of $\tilde{A}, \tilde{A}_y, \tilde{A}_{yy}, \tilde{A}_{gas}$ [19]):

$$\tilde{A}(y,p) = \int_0^\infty A(y,\tau) e^{-p\tau}\, d\tau, \tag{45}$$

then the problem for $A(y,\tau)$ is reduced to the boundary value problem for an ordinary differential equation

$$\tilde{A}_{yy} - \mathrm{Pr}\, p\, \tilde{A} = -\mathrm{Pr}\, A_0(y), \quad 0 < y < 1; \tag{46}$$

$$\tilde{A}(0,p) = \tilde{A}(p), \quad y = 0; \tag{47}$$

$$\tilde{A}_y + \mathrm{Bi}(\tilde{A} - \tilde{A}_{gas}) = 0, \quad y = 1. \tag{48}$$

The General solution of the equation (46) is as follows:

$$\tilde{A} = C_1 \,\mathrm{ch}\sqrt{\mathrm{Pr}\, p}\, y + C_2 \,\mathrm{sh}\sqrt{\mathrm{Pr}\, p}\, y + \frac{\sqrt{\mathrm{Pr}\, p}}{p} \int_0^y A_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\, p}\,(x-y)\right] dx; \tag{49}$$

with the constants C_1 and C_2, which are defined from boundary conditions (47), (48):

$$C_1 = \tilde{A}(p), \tag{50}$$

$$\begin{aligned} C_2 = \left[\sqrt{\mathrm{Pr}\, p}\,\mathrm{ch}\sqrt{\mathrm{Pr}\, p} + \mathrm{Bi}\,\mathrm{sh}\sqrt{\mathrm{Pr}\, p}\right]^{-1} \Bigg\{ \mathrm{Bi}\,\tilde{A}_{gas} - \tilde{A}(p)\left(\sqrt{\mathrm{Pr}\, p}\,\mathrm{sh}\sqrt{\mathrm{Pr}\, p} + \mathrm{Bi}\,\mathrm{ch}\sqrt{\mathrm{Pr}\, p}\right) - \\ - \frac{\mathrm{Bi}\,\sqrt{\mathrm{Pr}\, p}}{p}\mathrm{Pr} \int_0^1 A_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\, p}\,(x-1)\right] dx \Bigg\}. \end{aligned} \tag{51}$$

The original $A(y,\tau)$ is restored using the formula

$$A(y,\tau) = \frac{1}{2\pi i} \int\limits_{l-i\infty}^{l+i\infty} \tilde{A}(y,p) e^{p\tau}\, dp. \tag{52}$$

The integral (52) is taken along any straight line $Re\, p = l > s_0$, where s_0 is the growth index of the function $A(y,\tau)$, and is understood in the sense of the main value.

The task for determining the image $\tilde{B}(y,p)$ is exactly the same as the task (46)–(48) with the replacement of the right part: $-\mathrm{Pr}\, A_0(y)$ for $-\mathrm{Pr}\, B_0(y) - 4\tilde{A}$. Thus, this function is found by the formula:

$$\begin{aligned} \tilde{B} = C_3 \,\mathrm{ch}\sqrt{\mathrm{Pr}\, p}\, y + C_4 \,\mathrm{sh}\sqrt{\mathrm{Pr}\, p}\, y + \frac{\sqrt{\mathrm{Pr}\, p}}{p} \int_0^y B_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\, p}\,(x-y)\right] dx - \\ - \frac{2C_1 y}{\sqrt{\mathrm{Pr}\, p}}\,\mathrm{ch}\sqrt{\mathrm{Pr}\, p} - \frac{2C_2 y}{\sqrt{\mathrm{Pr}\, p}}\,\mathrm{sh}\sqrt{\mathrm{Pr}\, p} + \frac{2y}{p} \int_0^y A_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\, p}\,(x-y)\right] dx, \end{aligned} \tag{53}$$

with constants C_3 and C_4 defined from boundary conditions:

$$C_3 = \tilde{B}(p), \tag{54}$$

$$\begin{aligned} C_4 = \left[\sqrt{\mathrm{Pr}\, p}\,\mathrm{ch}\sqrt{\mathrm{Pr}\, p} + \mathrm{Bi}\,\mathrm{sh}\sqrt{\mathrm{Pr}\, p}\right]^{-1} \Bigg\{ \mathrm{Bi}\,\tilde{B}_{gas} - \tilde{B}(p)\left(\sqrt{\mathrm{Pr}\, p}\,\mathrm{sh}\sqrt{\mathrm{Pr}\, p} + \mathrm{Bi}\,\mathrm{ch}\sqrt{\mathrm{Pr}\, p}\right) + \\ + \mathrm{Pr} \int_0^1 B_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\, p}\,(x-1)\right] dx - \frac{2(1+\mathrm{Bi})}{p} \int_0^1 A_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\, p}\,(x-1)\right] dx + \\ + \frac{2\sqrt{\mathrm{Pr}\, p}}{p}\int_0^1 A_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\, p}\,(x-1)\right] dx + 2C_1\left(\frac{\mathrm{ch}\sqrt{\mathrm{Pr}\, p}}{\sqrt{\mathrm{Pr}\, p}} + \mathrm{sh}\sqrt{\mathrm{Pr}\, p} + \frac{\mathrm{Bi}\,\mathrm{ch}\sqrt{\mathrm{Pr}\, p}}{\sqrt{\mathrm{Pr}\, p}}\right) + \\ + 2C_2\left(\frac{\mathrm{sh}\sqrt{\mathrm{Pr}\, p}}{\sqrt{\mathrm{Pr}\, p}} + \mathrm{ch}\sqrt{\mathrm{Pr}\, p} + \frac{\mathrm{Bi}\,\mathrm{sh}\sqrt{\mathrm{Pr}\, p}}{\sqrt{\mathrm{Pr}\, p}}\right) \Bigg\}. \end{aligned} \tag{55}$$

You can show using the explicit formulas (49)–(51) and asymptotic representations: $\mathrm{sh}x \sim x + x^3/6,\ \mathrm{ch}x \sim 1 + x^2/2$ for $x \to 0$ [20], that

$$\lim_{\tau\to\infty} A(y,\tau) = \lim_{p\to 0} p\tilde{A}(y,p) = A^0(y),$$

where $A^0(y)$ is a stationary solution for the function $A(y,\tau)$ in (38). When proving, keep in mind that the functions $A_{gas}(\tau)$ and $A(\tau)$ are originals along with their first derivatives [19] and assume the existence of limits: $\lim\limits_{\tau\to\infty} A_{gas}(\tau) = \lim\limits_{p\to 0} p\tilde{A}_{gas}(p) = A^0_{gas}$, $\lim\limits_{\tau\to\infty} A(\tau) = \lim\limits_{p\to 0} p\tilde{A}(p) = A$. In addition, the condition (37) must be met.

Similarly, it is shown that

$$\lim_{\tau\to\infty} B(y,\tau) = \lim_{p\to 0} p\tilde{B}(y,p) = B^0(y),$$

that is, as time increases, the temperature perturbation becomes stationary, provided that the functions $B_{gas}(\tau)$ and $B(\tau)$ are originals along with their first derivatives and there are limits: $\lim\limits_{\tau\to\infty} B_{gas}(\tau) = \lim\limits_{p\to 0} p\tilde{B}_{gas}(p) = B^0_{gas}$, $\lim\limits_{\tau\to\infty} B(\tau) = \lim\limits_{p\to 0} p\tilde{B}(p) = B$.

Thus, the fair

Theorem 1. *Problem solving for the functions $A(y,\tau)$, $B(y,\tau)$ are determined by the inverse Laplace transform by the formulas* (49), (53)*, and with the growth of time, they reach a stationary regime, if $A_{gas}(\tau) \to A^0_{gas}, B_{gas}(\tau) \to B^0_{gas}$, $A(\tau) \to A, B(\tau) \to B$ when $\tau \to \infty$ and the condition* (37) *is met.*

5. Determination of the mixture concentration

Applying to the initial-boundary problem for the concentration the mixture of Laplace transform, obtain for the image $\tilde{H}(y,p)$ task

$$\tilde{H}_{yy} - \mathrm{Sc}p\,\tilde{H} = -\mathrm{Sc}H_0(y) + \mathrm{SrPr}\,A_0(y) - \mathrm{SrPr}p\,\tilde{A}, \quad 0 < y < 1; \tag{56}$$

$$\tilde{H}_y + \mathrm{Sr}\tilde{A}_y = 0, \quad y = 0; \tag{57}$$

$$\tilde{H}_y + \mathrm{Sr}\tilde{A}_y = 0, \quad y = 1. \tag{58}$$

The General solution of the equation (56) for $\mathrm{Pr} \neq \mathrm{Sc}$ is as follows:

$$\tilde{H} = C_5\,\mathrm{ch}\sqrt{\mathrm{Sc}\,p}\,y + C_6\,\mathrm{sh}\sqrt{\mathrm{Sc}\,p}\,y +$$

$$+\frac{1}{\sqrt{\mathrm{Sc}\,p}}\int_0^y (\mathrm{Sc}H_0(x) - \mathrm{SrPr}A_0(x))\mathrm{sh}\left[\sqrt{\mathrm{Sc}\,p}\,(x-y)\right]dx - \tag{59}$$

$$-\frac{\mathrm{SrPr}}{\mathrm{Pr}-\mathrm{Sc}}\left(C_1\mathrm{ch}\sqrt{\mathrm{Pr}\,p}\,y + C_2\mathrm{sh}\sqrt{\mathrm{Pr}\,p}\,y + \frac{\sqrt{\mathrm{Pr}\,p}}{p}\int_0^y A_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\,p}\,(x-y)\right]dx\right),$$

with constants C_5 and C_6 defined from boundary conditions (57), (58):

$$C_6 = \frac{C_2\,\mathrm{Sr}\,\sqrt{\mathrm{PrSc}}}{\mathrm{Pr}-\mathrm{Sc}}, \tag{60}$$

$$C_5 = \left[\sqrt{\mathrm{Sc}\,p}\,\mathrm{sh}\sqrt{\mathrm{Sc}\,p}\right]^{-1}\left\{\int_0^1 (\mathrm{Sc}H_0(x) - \mathrm{SrPr}A_0(x))\mathrm{ch}\left[\sqrt{\mathrm{Sc}\,p}\,(x-1)\right]dx + \right.$$
$$+\frac{\mathrm{SrSc}}{\mathrm{Pr}-\mathrm{Sc}}\left[\sqrt{\mathrm{Pr}\,p}\left(C_1\,\mathrm{sh}\sqrt{\mathrm{Pr}\,p} + C_2\,\mathrm{ch}\sqrt{\mathrm{Pr}\,p}\right) - \right. \tag{61}$$
$$\left.\left. -\mathrm{Pr}\int_0^1 A_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\,p}\,(x-1)\right]dx\right\} - \frac{C_2\,\mathrm{Sr}\sqrt{\mathrm{PrSc}}}{\mathrm{Pr}-\mathrm{Sc}}\,\mathrm{cth}\sqrt{\mathrm{Sc}\,p}\,.$$

The task for defining an image $\tilde{G}(y,p)$ is exactly the same as the task (56)–(58) with replacing the right part: $-\mathrm{Sc}H_0(y) + \mathrm{SrPr}\,A_0(y) - \mathrm{SrPr}p\,\tilde{A}$ for $-\mathrm{Sc}G_0(y) + \mathrm{SrPr}\,B_0(y) - \mathrm{SrPr}p\,\tilde{B} - 4\tilde{H}$.

The General solution for $\tilde{G}(y,p)$ when $\mathrm{Pr} \neq \mathrm{Sc}$ has the form:

$$\tilde{G} = C_7\,\mathrm{ch}\sqrt{\mathrm{Sc}\,p}\,y + C_8\,\mathrm{sh}\sqrt{\mathrm{Sc}\,p}\,y +$$
$$+\frac{1}{\sqrt{\mathrm{Sc}\,p}}\int_0^y (\mathrm{Sc}G_0(x) - \mathrm{SrPr}B_0(x))\mathrm{sh}\left[\sqrt{\mathrm{Sc}\,p}\,(x-y)\right]dx -$$
$$-\frac{2C_5 y}{\sqrt{\mathrm{Sc}\,p}}\,\mathrm{ch}\sqrt{\mathrm{Sc}\,p}\,y - \frac{2C_6 y}{\sqrt{\mathrm{Sc}\,p}}\,\mathrm{sh}\sqrt{\mathrm{Sc}\,p}\,y -$$
$$-\frac{\mathrm{SrPr}}{\mathrm{Pr}-\mathrm{Sc}}\left(C_3\,\mathrm{ch}\sqrt{\mathrm{Pr}\,p}\,y + C_4\,\mathrm{sh}\sqrt{\mathrm{Pr}\,p}\,y + \frac{\sqrt{\mathrm{Pr}\,p}}{p}\int_0^y B_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\,p}\,(x-y)\right]dx - \right. \tag{62}$$
$$\left. -\frac{2C_1 y}{\sqrt{\mathrm{Pr}\,p}}\,\mathrm{ch}\sqrt{\mathrm{Pr}\,p}\,y - \frac{2C_2 y}{\sqrt{\mathrm{Pr}\,p}}\,\mathrm{sh}\sqrt{\mathrm{Pr}\,p}\,y + \frac{2y}{p}\int_0^y A_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\,p}\,(x-y)\right]dx\right) +$$
$$+\frac{2y}{\mathrm{Sc}p}\int_0^y (\mathrm{Sc}H_0(x) - \mathrm{SrPr}A_0(x))\mathrm{ch}\left[\sqrt{\mathrm{Sc}\,p}\,(x-y)\right]dx,$$

where the constants C_7 and C_8 are defined from the boundary conditions as follows:

$$C_8 = \frac{2C_5}{\mathrm{Sc}p} + \frac{\mathrm{Sr}\,\mathrm{Sc}\,(\mathrm{Pr}p\,C_4 - 2C_1)}{\sqrt{\mathrm{PrSc}}\,p\,(\mathrm{Pr}-\mathrm{Sc})}, \tag{63}$$

$$C_7 = \left[\sqrt{\mathrm{Sc}\,p}\,\mathrm{sh}\sqrt{\mathrm{Sc}\,p}\right]^{-1}\left\{2C_5\left(\frac{\mathrm{ch}\sqrt{\mathrm{Sc}\,p}}{\sqrt{\mathrm{Sc}\,p}} + \mathrm{sh}\sqrt{\mathrm{Sc}\,p}\right) + 2C_6\left(\frac{\mathrm{sh}\sqrt{\mathrm{Sc}\,p}}{\sqrt{\mathrm{Sc}\,p}} + \mathrm{ch}\sqrt{\mathrm{Sc}\,p}\right) + \right.$$
$$+\int_0^1 (\mathrm{Sc}G_0(x) - \mathrm{SrPr}B_0(x))\mathrm{ch}\left[\sqrt{\mathrm{Sc}\,p}\,(x-1)\right]dx +$$
$$+\frac{\mathrm{SrSc}}{\mathrm{Pr}-\mathrm{Sc}}\left[\sqrt{\mathrm{Pr}\,p}\left(C_3\,\mathrm{sh}\sqrt{\mathrm{Pr}\,p} + C_4\,\mathrm{ch}\sqrt{\mathrm{Pr}\,p}\right) - \mathrm{Pr}\int_0^1 B_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\,p}\,(x-1)\right]dx - \right.$$
$$-2C_1\left(\frac{\mathrm{ch}\sqrt{\mathrm{Pr}\,p}}{\sqrt{\mathrm{Pr}\,p}} + \mathrm{sh}\sqrt{\mathrm{Pr}\,p}\right) - 2C_2\left(\frac{\mathrm{sh}\sqrt{\mathrm{Pr}\,p}}{\sqrt{\mathrm{Sc}\,p}} + \mathrm{ch}\sqrt{\mathrm{Pr}\,p}\right) +$$
$$\left. + \frac{2}{p}\int_0^1 A_0(x)\mathrm{ch}\left[\sqrt{\mathrm{Pr}\,p}\,(x-1)\right]dx - \frac{2\sqrt{\mathrm{Pr}\,p}}{p}\int_0^1 A_0(x)\mathrm{sh}\left[\sqrt{\mathrm{Pr}\,p}\,(x-1)\right]dx\right] -$$
$$-\frac{2}{\mathrm{Sc}p}\int_0^1 (\mathrm{Sc}H_0(x) - \mathrm{SrPr}A_0(x))\mathrm{ch}\left[\sqrt{\mathrm{Sc}\,p}\,(x-1)\right]dx +$$
$$\left. + \frac{2}{\sqrt{\mathrm{Sc}\,p}}\int_0^1 (\mathrm{Sc}H_0(x) - \mathrm{SrPr}A_0(x))\mathrm{sh}\left[\sqrt{\mathrm{Sc}\,p}\,(x-1)\right]dx\right\} - C_8\,\mathrm{cth}\sqrt{\mathrm{Sc}\,p}\,. \tag{64}$$

You can show using the formulas (59)–(61) that

$$\lim_{\tau\to\infty} H(y,\tau) = \lim_{p\to 0} p\tilde{H}(y,p) = H^0(y),$$

where $H^0(y)$ is a stationary solution for the function $H(y,\tau)$ in (39). When output, you must again assume that there are limits: $\lim\limits_{\tau\to\infty} A_{gas}(\tau) = A^0_{gas}$, $\lim\limits_{\tau\to\infty} A(\tau) = A$. In addition, the condition (37) must be met.

Similarly, it is shown that

$$\lim_{\tau\to\infty} G(y,\tau) = \lim_{p\to 0} p\tilde{G}(y,p) = G^0(y),$$

where $G^0(y)$ is a stationary distribution for the function $G(y,\tau)$.

Thus, the fair

Theorem 2. *Problem solving for the functions $H(y,\tau)$, $G(y,\tau)$ are determined by the inverse Laplace transform by the formulas* (59), (62)*, and with the growth of time, they reach a stationary regime, if $A_{gas}(\tau) \to A^0_{gas}, B_{gas}(\tau) \to B^0_{gas}$, $A(\tau) \to A, B(\tau) \to B$ when $\tau \to \infty$ and the condition* (37) *is met.*

6. Determination of the velocity field

Applying the Laplace transform to a problem for speed reduces it to a boundary value problem for an ordinary differential equation

$$\tilde{U}_{yy} - p\tilde{U} = -U_0(r) - \tilde{F}(p), \quad 0 < y < 1; \tag{65}$$

$$\tilde{U}(0,p) = 0, \quad y = 0; \tag{66}$$

$$\int_0^1 \tilde{U}dy = 0, \quad y = 1; \tag{67}$$

$$\tilde{U}_y = -2(\tilde{A} + \mathrm{M}\tilde{H}), \quad y = 1. \tag{68}$$

The General solution of the equation (65) is written as follows:

$$\tilde{U} = C_9 \operatorname{ch}\sqrt{p}\, y + C_{10} \operatorname{sh}\sqrt{p}\, y + \frac{1}{\sqrt{p}} \int_0^y U_0(x) \operatorname{sh}\Big[\sqrt{p}\,(x-y)\Big]\, dx, \tag{69}$$

with constants C_9 and C_{10} defined from boundary conditions (66)–(68):

$$C_9 = -\frac{\tilde{F}(p)}{p}, \tag{70}$$

$$C_{10} = \frac{2(\tilde{A} + \mathrm{M}\tilde{H}) + \sqrt{p}\int\limits_0^1 U_0(x)\operatorname{ch}\sqrt{p}\,(x-1)\,dx + \tilde{F}(p)\operatorname{sh}\sqrt{p}}{p\operatorname{ch}\sqrt{p}}, \tag{71}$$

where the functions $\tilde{A}(y,p), \tilde{H}(y,p)$ are given by the formulas (49), (59) for $y = 1$, and the pressure gradient $\tilde{F}(p)$ is as follows:

$$\begin{gathered}\tilde{F}(p) = \Big[\sqrt{p}\operatorname{ch}\sqrt{p} - \operatorname{sh}\sqrt{p}\Big]^{-1}\Big\{\sqrt{p}\,(\operatorname{ch}\sqrt{p} - 1)(2\tilde{A} + 2\mathrm{M}\tilde{H} - \\ - \int_0^1 U_0(x)\operatorname{ch}\sqrt{p}\,(x-1)\,dx) - p\operatorname{ch}\sqrt{p}\int_0^1\Big[\int_0^y U_0(x)\operatorname{ch}\sqrt{p}\,(x-1)\,dx\Big]dy\Big\}.\end{gathered} \tag{72}$$

You can derive equality from the expressions (69)–(72):

$$\lim_{p\to 0} p\widetilde{U}(r,p) = U^0(y), \tag{73}$$

where $U^0(y)$ is a stationary velocity distribution from (40). When you output (73), you must assume the existence of the limits: $\lim\limits_{\tau\to\infty} A_{gas}(\tau) = A^0_{gas}$, $\lim\limits_{\tau\to\infty} B_{gas}(\tau) = B^0_{gas}$, $\lim\limits_{\tau\to\infty} A(\tau) = A$, $\lim\limits_{\tau\to\infty} B(\tau) = B$ and the fulfillment of the condition (37).

Thus, the fair

Theorem 3. *Problem solving for the function $U(y,\tau)$ is determined by the inverse Laplace transform by the formulas* (69), (71)*, and with the growth of time, they reach a stationary regime, if $A_{gas}(\tau) \to A^0_{gas}, B_{gas}(\tau) \to B^0_{gas}$, $A(\tau) \to A, B(\tau) \to B$ when $\tau \to \infty$ and the condition* (37) *is met.*

7. Numerical solution

The obtained formulas in the Laplace images were used to find the temperature, concentration, and velocity fields of the mixture under certain conditions imposed on the external temperature $A_{gas}(\tau)$ and the solid wall temperature $A(\tau)$. In this purpose, the numerical method of the inverse Laplace transform was used using the quadrature formula of the highest degree of accuracy, constructed for the Riemann–Mellin integral [21]:

$$f(t) = \frac{1}{2\pi i} \int\limits_{c-i\infty}^{c+i\infty} F(\sigma) e^{\sigma t}\, d\sigma. \tag{74}$$

Let the image function $F(\sigma)$ is regular in the half-plane $Re\,\sigma > \alpha$. Replacing $\sigma = p/t + \alpha$ converts (74) to an integral

$$f(t) = \frac{1}{2\pi i}\, \frac{e^{\alpha t}}{t} \int\limits_{\varepsilon-i\infty}^{\varepsilon+i\infty} F^*(p) e^p\, dp, \tag{75}$$

here ε — any small positive number, and $F^*(p) = F(p/t + \alpha) = F(\sigma)$. It is assumed that $F^*(p)$ has the form $F^*(p) = \varphi(p)/p^k$, here $k > 0$, $\varphi(p)$ is regular in the half-plane $Re\,p > 0$ and there is $\lim\limits_{t\to\infty} \varphi(p) \neq 0; \infty$. Then the quadrature formula of the highest degree of accuracy is applied to the integral

$$\frac{1}{2\pi i} \int\limits_{\varepsilon-i\infty}^{\varepsilon+i\infty} \varphi(p) \frac{e^p}{p^k}\, dp$$

which has the form

$$\frac{1}{2\pi i} \int\limits_{\varepsilon-i\infty}^{\varepsilon+i\infty} \varphi(p) \frac{e^p}{p^k}\, dp \simeq \sum_{m=1}^{n} A_m \varphi(p_m), \tag{76}$$

and since

$$\varphi(p_m) = p_m^k F^*(p_m) = p_m^k\, F(p_m/t + \alpha),$$

then

$$f(t) \simeq \frac{e^{\alpha t}}{t} \sum_{m=1}^{n} A_m\, p_m^k\, F(p_m/t + \alpha), \tag{77}$$

moreover, the coefficients A_m and p_m nodes depend on k и n. The formula (77) was the basis of a program that performs the inverse Laplace transform. The coefficients A_m and the nodes p_m were taken from [22].

Using the numerical method, quantitative results were obtained for a model system with the following parameter values: $A^0_{gas} = 0.2$, $A = -0.1$, Sr $= 3$, Bi $= 2$, Pr $= 2$, Sc $= 1$, M $= 100$, $A_{gas}(\tau) = A^0_{gas} + \exp(-\lambda\tau)\sin(\omega\tau)$, $A(\tau) = A + \exp(-\lambda\tau)\sin(\omega\tau)$, here $\omega = 1$. Fig. 2–7 shows the evolution of dimensionless profiles of temperature, concentration, and velocity of the mixture at different times.

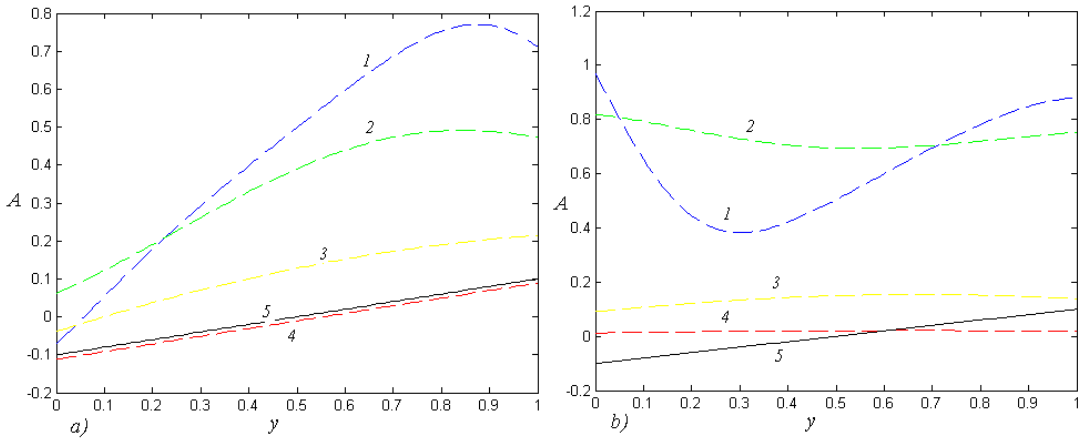

Fig. 2. The temperature profile at $\lambda = 1$: 1 — $\tau = 0.02$, 2 — $\tau = 0.2$, 3 — $\tau = 2.4$, 4 — $\tau = 4.5$, 5 — the stationary solution

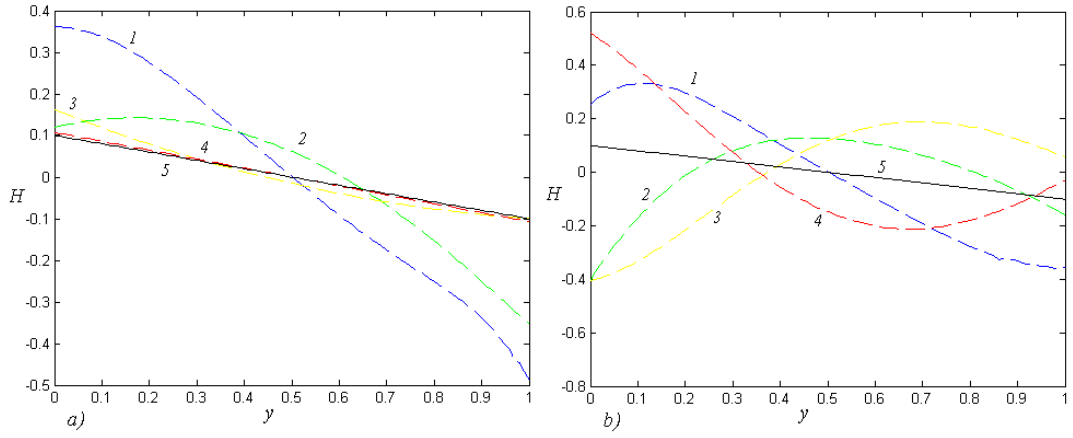

Fig. 3. The concentration profile at $\lambda = 1$: 1 — $\tau = 0.02$, 2 — $\tau = 0.2$, 3 — $\tau = 1.7$, 4 — $\tau = 4.8$, 5 — the stationary solution

If the functions $A_{gas}(\tau), A(\tau)$ have finite limits at $\tau \to \infty$, equal to A^0_{gas} and A, respectively, and the condition (37) is met, then there is convergence to the stationary distribution (see Fig. $2a, 3a, 4a$ at $\lambda = 1$). If these functions have no limits at $\tau \to \infty$ (either the limits exist, but the connection between A^0_{gas} and A is broken), then non-stationary solutions do not converge to stationary solutions with increasing time (see Fig. 2b, 3b, 4b at $\lambda = 1$).

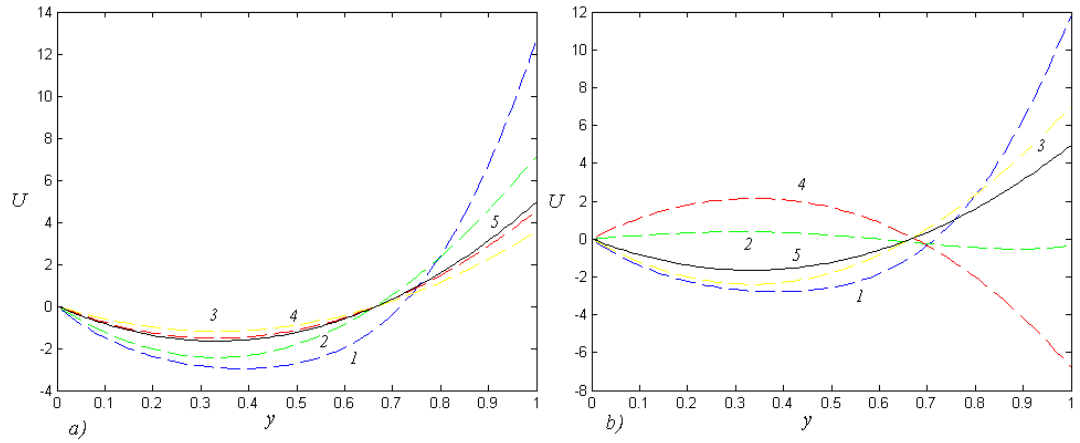

Fig. 4. The velocity profile at $\lambda = 1$: 1 — $\tau = 0.04$, 2 — $\tau = 1.0$, 3 — $\tau = 1.4$, 4 — $\tau = 3.14$, 5 — the stationary solution

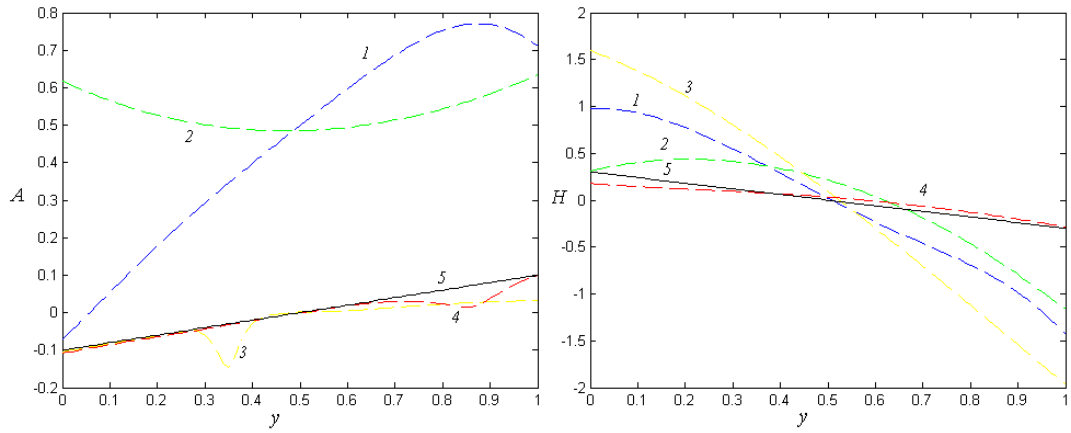

Fig. 5. The temperature and concentration profiles at $\lambda = 10^{-3}$: 1 — $\tau = 0.03$, 2 — $\tau = 0.3$, 3 — $\tau = 35.6$, 4 — $\tau = 37.85$, 5 — the stationary solution

For Fig. 5, 6 presents temperature, concentration, and velocity profiles for $\lambda = 10^{-3}$. It takes a longer period of time for the solution to return to the steady state, and there are fluctuations. The dependence of the speed $U(y, \tau)$ on the parameter M was also studied (see Fig. 7). It turned out that the non-stationary solution quickly switches to the stationary regime for any M.

Analyzing the numerical solution for the function $U(y, \tau)$, we conclude that she takes a minimum value for $y = 1/3$, as well as $U < 0$ for $0 < y < 2/3$ and $U > 0$ for $2/3 < y < 1$, which corresponds to the result obtained in the formula (40). It follows that the current changes direction at a depth equal to 2/3 of the thickness of the liquid layer.

Fig. 8 shows the trajectories of liquid particles (current lines) and the surface of the current when moving a viscous heat-conducting binary mixture with a flat free boundary. It can be seen that there is a return rotationally-symmetric flow of the liquid, which occurs under the influence of a pressure gradient. The resulting motion is a vortex in the ry plane with the center shifted to the free boundary. In this case, the maximum speed is achieved on a free surface.

Let's see what happens to the rest of the required functions. As a result of heat exposure,

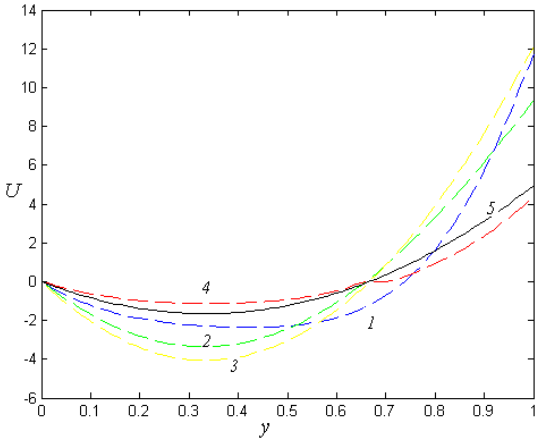

Fig. 6. The velocity profile at $\lambda = 10^{-3}$: 1 — $\tau = 0.5$, 2 — $\tau = 4.5$, 3 — $\tau = 35.2$, 4 — $\tau = 37.8$, 5 — the stationary solution

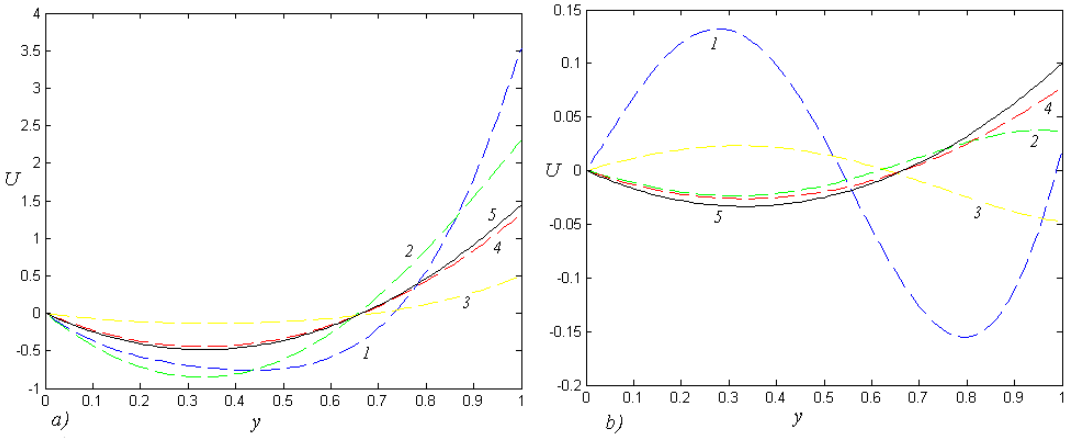

Fig. 7. The velocity profile for different values of the parameter M: *a*) M = 10, *b*) M = 1

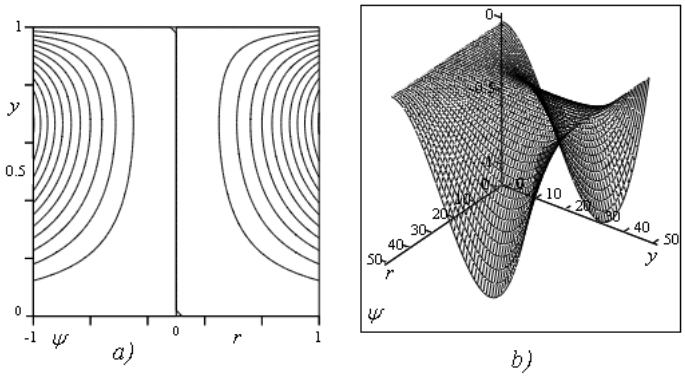

Fig. 8. *a*) the trajectories of liquid particles, *b*) the surface current

the temperature $A(y,\tau)$ increases and the concentration $H(y,\tau)$ decreases. There is a thermodiffusion effect-the Soret effect. Anomalous thermodiffusion occurs, in which light components tend to move to colder areas, and heavy components end up in areas with increased temperature (since c in the system (1) is the concentration of the light component).

Conclusion

Rotationally-symmetric motion of a binary mixture with a flat free boundary at small Marangoni numbers is investigated. The problem is reduced to the inverse linear initial-boundary value problem for parabolic equations. Using Laplace transformation properties the exact analytical solution is obtained. It is shown that a stationary solution is the limiting one with the growth of time if there is a certain relationship between the temperature of the solid wall and the external temperature of the gas. If there is no connection, the convergence to the stationary solution is broken. Some examples of numerical reconstruction of the temperature, concentration and velocity fields are given, which confirm the theoretical conclusions.

References

[1] N.Darabi, Two-dimensional motion of a binary mixture such as Hiemenz in a flat layer, *Journal of the Siberian Federal University. Mathematics and physics*, **8**(2015), no. 3, 260–272.

[2] R.V.Birikh, *J. Appl. Mech. Tech. Phys.*, **7**(1966), 43–44. DOI: 10.1007/BF00914697

[3] A.G.Kirdyashkin, Thermogravitational and thermocapillary convection in a horizontal fluid layer, In: Fluid mechanics and transport processes in zero gravity, Sverdlovsk, UNC of the USSR Academy of Sciences, 1983, 81–86 (in Russian).

[4] A.G.Kirdyashkin, *International Journal of Heat and Mass Transfer*, **27**(1984), no. 8, 1205–1218. DOI: 10.1016/0017-9310(84)90048-6

[5] A.G.Kirdyashkin, Thermocapillary periodic flows, Preprint no. 8. Novosibirsk: IGiG SO AS USSR, 1985 (in Russian).

[6] A.F.Sidorov, *J. Appl. Mech. Tech. Phys.*, **30**(1989), 197–203. DOI: 10.1007/BF00852164

[7] S.N.Aristov, K.G.Shvartz, Vortex flows of advective nature in a rotating liquid layer, Perm, PSU publishing House, 2006 (in Russian).

[8] O.N.Goncharova, O.A.Kabov, *Microgravity Sci. Thechnol.*, **21**(2009), 129–137. DOI: 10.1007/s12217-009-9108-x

[9] V.K.Andreev, Birikh's Solution of convection equations and some of its generalizations, Preprint no. 1-10, Krasnoyarsk, Institute of Computational Modeling SB RAS, 2010 (in Russian).

[10] S.N.Aristov, K.G.Schwartz, Vortex flows in thin layers of liquid, Kirov, VyatSU, 2011 (in Russian).

[11] V.V.Pukhnachev, Nonstationary analogs of the Birikh solution, *The Bulletin of Irkutsk State University. Series Mathematics*, **3**(2011), no. 1, 61–69 (in Russian).

[12] V.K.Andreev, V.B.Bekezhanova, *J. Appl. Mech. Tech. Phy.*, **54**(2013), 171–184. DOI: 10.1134/S0021894413020016

[13] S.N.Aristov, E.Yu.Prosviryakov, On a class of analytical solutions for stationary axisymmetric Benard-Marangoni convection of a viscous incompressible fluid, *J. Samara State Tech. Univ., Ser. Phys. Math. Sci.*, **32**(2013), no. 3, 110–118 (in Russian).

[14] A.V.Gorshkov, E.Yu.Prosviryakov, *Computer research and modeling*, **8**(2016), no. 6, 927–940 (in Russian). DOI: 10.20537/2076-7633-2016-8-6-927-940

[15] V.V.Privalova, E.Yu.Prosviryakov, Couette–Hiemenz exact solutions for the steady creeping convective flow of a viscous incompressible fluid with allowance made for heat recovery *J. Samara State Tech. Univ., Ser. Phys. Math. Sci.*, **22**(2018), no. 3, 532–548.

[16] K.Hiemenz, Die Grenzschicht an einem in den gleichformigen Flussigkeitsstrom eingetauchten geraden Kreiszylinder, *Dinglers Politech. J.*, **326**(1911), 321.

[17] V.K.Andreev, V.E.Zachvataev, E.A.Ryabitskiy, Thermocapillary instability, Novosibirsk, Nauka, 2000 (in Russian).

[18] L.V.Ovsyannikov, Group analysis of differential equations, Moscow, Nauka, 1978 (in Russian).

[19] M.A.Lavrentiev, B.V.Shabat, Methods of the theory of functions of a complex variable, Moscow, Nauka, 1973 (in Russian).

[20] M.Abramovits, I.Stigan, Handbook of special functions, Moscow, Nauka, 1979 (in Russian).

[21] V.I.Krylov, N.S.Skoblya, Reference book on the numerical inversion of the Laplace transform, Minsk, Science and technology, 1968 (in Russian).

[22] A.P.Prudnikov, V.A.Ditkin, Operational calculus, Moscow, Higher school, 1975 (in Russian).

Вращательно-осесимметричное движение бинарной смеси с плоской свободной границей при малых числах Марангони

Виктор К. Андреев
Институт вычислительного моделирования СО РАН
Красноярск, Российская Федерация
Сибирский федеральный университет
Российская Федерация
Наталья Л. Собачкина
Сибирский федеральный университет
Российская Федерация

Аннотация. Исследовано вращательно-симметричное движение плоского слоя бинарной смеси со свободной границей при малых числах Марангони. Задача сводится к обратной линейной начально-краевой задаче для параболических уравнений. В изображениях по Лапласу получено точное аналитическое решение. Найдено стационарное решение задачи и доказано, что оно является предельным с ростом времени при условии существования определенной связи между температурой твердой стенки и внешней температурой газа. В случае отсутствия связи сходимость к стационарному решению нарушается. Приведены примеры численного восстановления полей температуры, концентрации и скорости, подтверждающие теоретические выводы.

Ключевые слова: бинарная смесь, свободная граница, обратная задача, градиент давления, стационарное решение, преобразование Лапласа, тепловое число Марангони.

DOI: 10.17516/1997-1397-2020-13-2-213-217
УДК 539.374

Anisotropic Antiplane Elastoplastic Problem

Sergei I. Senashov*
Irina L. Savostyanova†
Reshetnev Siberian State University of Science and Technology
Krasnoyarsk, Russian Federation
Olga N. Cherepanova‡
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 10.11.2019, received in revised form 11.01.2020, accepted 20.02.2020

Abstract. In this work we solve an anisotropic antiplane elastoplastic problem about stress state in a body weakened by a hole bounded by a piecewise-smooth contour. We give the conservation laws which allowed us to reduce calculations of stress components to a contour integral over the contour of the hole. The conservation laws allowed us to find the boundary between the elastic and plastic areas.

Keywords: anisotropic elastoplastic problem, antiplane stress state, conservation laws.

Citation: S.I.Senashov, I.L.Savostyanova, O.N.Cherepanova, Anisotropic Antiplane Elastoplastic Problem, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 213–217.
DOI: 10.17516/1997-1397-2020-13-2-213-217.

Introduction

Fields of shifts and stresses in the case under consideration are the following [1]

$$u = v = 0, \quad w = w(x,y) \quad \sigma_x = \sigma_y = \sigma_z = \tau_{xy} = 0, \quad \tau_{xz} = \tau^1(x,y), \quad \tau_{yz} = \tau^2(x,y). \tag{1}$$

Here u, v, w are shift vector components, $\sigma_x, \sigma_y, \sigma_z, \tau_{xy}, \tau_{xz}, \tau_{yz}$ are stress components, x, y, z the Cartesian coordinates, axis directed parallel to the element.

In the elastic zone there are the relations

$$\frac{\partial \tau^1}{\partial x} + \frac{\partial \tau^2}{\partial y} = 0 \ (equilibrium\ equation), \tag{2}$$

$$\tau^1 = \frac{\partial w}{\partial x}, \tau^2 = G_2 \frac{\partial w}{\partial y} \ (Hooke's\ law). \tag{3}$$

Here G_i are constants called elastic moduli [2].

From (2), (3) there arise relations in the elastic zone

$$G_1 \frac{\partial^2 w}{\partial x^2} + G_2 \frac{\partial^2 w}{\partial y^2} = 0, \tag{4}$$

$$G_2 \frac{\partial \tau^1}{\partial y} = G_1 \frac{\partial \tau^2}{\partial x}. \tag{5}$$

*sen@mail.sibsau.ru https://orcid.org/0000-0001-5542-4781
†ruppa@inbox.ru https://orcid.org/0000-0002-9675-7109
‡cheronik@mail.ru

From (2) and (5) it follows that τ^1, τ^2 satisfy the system of linear equations

$$F_1 = \frac{\partial \tau^1}{\partial x} + \frac{\partial \tau^2}{\partial y} = 0, \quad F_2 = \frac{\partial \tau^1}{\partial y} - n\frac{\partial \tau^2}{\partial x} = 0, \tag{6}$$

where $n = G_1/G_2$.

In the plastic zone there holds the relation (2), and also

$$a_{13}(\tau^1)^2 + a_{23}(\tau^2)^2 = 1 \ (yield\ condition), \tag{7}$$

$$\tau^2 \frac{\partial w}{\partial x} = \tau^1 \frac{\partial w}{\partial y} \ (Hencky's\ equation). \tag{8}$$

Here a_{13}, a_{23} are constants called anisotropy coefficients.

On the boundary of the elastic and plastic areas the stresses and shifts are supposed to be continuous.

1. Conservation laws

By a conservation law for the system of equations (6) we shall call the relation of the form of

$$\frac{\partial A\left(x, y, \tau^1, \tau^2\right)}{\partial x} + \frac{\partial B\left(x, y, \tau^1, \tau^2\right)}{\partial y} = \omega^1 F_1 + \omega^2 F_2, \tag{9}$$

where $\omega^i = \omega^i(x, y, \tau^1, \tau^2)$ are some functions not identically zero simultaneously.

Note. A more general definition of conservation laws and their use in mechanics of a solid body being deformed can be studied for example in [3–5].

For the purposes that are set in this article a simplified formulation in the form of (9) will suit fine.

In (9) the values A, B are called conserved current components.

Let us assume that the components A, B appear as follows

$$A = \alpha^1\tau^1 + \beta^1\tau^2 + \gamma^1, \quad B = \alpha^2\tau^1 + \beta^2\tau^2 + \gamma^2, \tag{10}$$

where $\alpha^i = \alpha^i(x, y)$, $\beta^i = \beta^i(x, y)$, $\gamma^i = \gamma^i(x, y)$ are some smooth functions to be determined.

Let us substitute (10) into (9), as a result we obtain

$$\begin{aligned} &\alpha^1_x\tau^1 + \alpha^1\tau^1_x + \beta^1_x\tau^2 + \beta^1\tau^2_x + \gamma^1_x + \alpha^2_y\tau^1 + \alpha^2\tau^1_y + \beta^2_y\tau^2 + \beta^2\tau^2_y + \gamma^2_y = \\ &= \omega^1\left(\tau^1_x + \tau^2_y\right) + \omega^2\left(\tau^1_y - n\tau^2_x\right) = 0, \end{aligned} \tag{11}$$

where the index below stands for a derivative with respect to the corresponding variable.

From (11) we obtain

$$\alpha^1 = \omega^1, \quad \beta^1 = -n\omega^2, \quad \alpha^2 = \omega^2, \quad \beta^2 = \omega^1, \quad \alpha^1_x + \alpha^2_y = 0, \quad \beta^1_x + \beta^2_y = 0, \quad \gamma^1_x + \gamma^2_y = 0. \tag{12}$$

From (12) excluding ω^i we obtain

$$\alpha^1 = \beta^2, \quad \beta^1 = -n\alpha^2, \quad \alpha^1_x - n\beta^1_y = 0, \quad \beta^1_x + \alpha^1_y = 0, \quad \gamma^1_x + \gamma^2_y = 0. \tag{13}$$

By virtue of relations (12) the conserved current components are written as

$$A = \alpha^1\tau^1 + \beta^1\tau^2 + \gamma^1, \quad B = \frac{-\beta^1}{n}\tau^1 + \alpha^1\tau^2 + \gamma^2. \tag{14}$$

Since the right-hand part (9) is equal to zero, according to Green's formula we obtain

$$\begin{aligned}&\iint_S (A_x + B_y)\,dxdy = \oint_{\partial S} Ady - Bdx = \\ &= \oint_{\partial S} (\alpha^1\tau^1 + \beta^1\tau^2 + \gamma^1)dy - \left(\frac{-\beta^1}{n}\tau^1 + \alpha^1\tau^2 + \gamma^2\right)dx = 0,\end{aligned} \tag{15}$$

where S is the area, ∂S is its piecewise-smooth boundary. All the functions in (15) are supposed to be smooth.

2. Elastoplastic problem for an arbitrary hole in case when the plastic area surrounds the entire hole

Assume C is a piecewise-smooth contour, there is a load applied to it

$$l_1\tau^1 + l_2\tau^2 = \tau_n, \quad |\tau_n| \leqslant \sqrt{\frac{l_1^2 a_{23} + l_2^2 a_{13}}{a_{13}a_{23}}}, \tag{16}$$

where (l_1, l_2) are normal's vector components to contour C. The plastic area's contour L surrounds entirely the hole C. See Fig. 1.

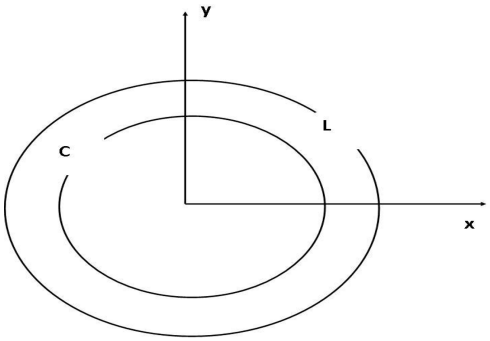

Fig. 1. Elastic-plastic border near the hole C

In this case on contour C, apart from the condition (16), also fulfilled is the yield condition (7). Thus on C there are two conditions:

$$l_1\tau^1 + l_2\tau^2 = \tau_n = \tau_n, \quad a_{13}(\tau^1)^2 + a_{23}(\tau^2)^2 = 1. \tag{17}$$

From the conditions (17) we find the stress components on contour C:

$$\tau^1 = -\frac{l_2}{l_1}\tau^2 + \frac{\tau_n}{l_1}, \quad \tau^2 = \frac{a_{13}l_2\tau_n \mp l_1\sqrt{l_1^2 a_{23} + l_2^2 a_{13} - a_{13}a_{23}\tau_n^2}}{l_1^2 a_{23} + l_2^2 a_{13}}. \tag{18}$$

From this point on, to be definite, in formulas (18) we will be selecting the upper sign.

3. The use of conservation laws to find stress components in the area

Assume the point $M\,(x_m, y_m)$ lies beyond the contour C. Let us draw a circumference with radius ε with the centre at the point M. We have $\varepsilon : (x - x_m)^2 + (y - y_m)^2 = \varepsilon^2$. Assume D is

a line connecting the point M with the contour C. We obtain a closed contour consisting of the circumference ε, the segmant P and the contour C. See Fig. 2.

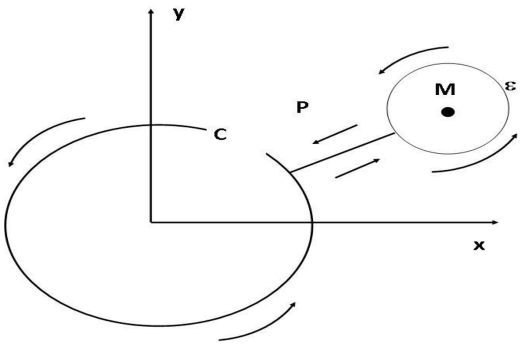

Fig. 2. Calculating the contour integral around the singular point M

From (15) we obtain

$$\oint_C Ady - Bdx + \int_{P^+} Ady - Bdx + \int_{P^-} Ady - Bdx + \oint_\varepsilon Ady - Bdx = 0. \tag{19}$$

The sum of the second and the third summands in (19) is equal to zero, because the integrals are calculated in different directions. Finally from (19) we have

$$\int_C Ady - Bdx = -\oint_\varepsilon Ady - Bdx. \tag{20}$$

Let us convert the right-hand part of equation (20) introducing parametrisation $x = \varepsilon\cos t$, $y = \varepsilon\sin t$, $0 \leqslant t \leqslant 2\pi$. As a result we have

$$\oint_\varepsilon Ady - Bdx = \varepsilon\int_0^{2\pi} (A\cos t + B\sin t)\, dt. \tag{21}$$

Assume in (15)

$$\alpha^1 = \frac{x}{x^2 + ny^2}, \beta^1 = -\frac{y}{x^2 + ny^2} \tag{22}$$

Then from (21) we obtain

$$\oint_\varepsilon A_1 dy - B_1 dx = \varepsilon\int_0^{2\pi} (A_1\cos t + B_1\sin t)\, dt = \int_0^{2\pi} \tau^1 dt = 2\pi\tau^1\,(x_m, y_m). \tag{23}$$

The last equality in (23) is obtained with the use of the mean-value theorem with ε tending to zero.

Assume in (15)

$$\alpha^1 = \frac{\sqrt{n}y}{x^2 + ny^2}, \quad \beta^1 = \frac{1}{\sqrt{n}}\frac{x}{x^2 + ny^2}. \tag{24}$$

Then from (21) we obtain

$$\oint_\varepsilon A_2 dy - B_2 dx = \varepsilon\int_0^{2\pi} (A_2\cos t + B_2\sin t)\, dt = \int_0^{2\pi} \tau^2 dt = 2\pi\tau^2\,(x_m, y_m). \tag{25}$$

The last equality in (25) is obtained with the use of the mean-value theorem with ε tending to zero.

From formula (20), and also from (23) and (25) we obtain

$$\int_C A_1 dy - B_1 dx = -2\pi\tau^1(x_m, y_m), \quad \int_C A_2 dy - B_2 dx = -2\pi\tau^2(x_m, y_m). \tag{26}$$

Conclusion

Formulas (26) offer the opportunity to find stress components in any point x_m, y_m beyond the contour C. This allows us to determine the boundary between the elastic and plastic areas. If the plasticity condition is met $a_{13}(\tau^1)^2 + a_{23}(\tau^2)^2 = 1$ at the point x_m, y_m then this point belongs to the plastic area, if in the point the condition $a_{13}(\tau^1)^2 + a_{23}(\tau^2)^2 < 1$ is met, then to the elastic area.

Note. The formulas found above allow us to solve elastoplastic problems even if the plastic contour does not entirely surrounds the contour C, provided that on the contour C the plasticity condition (7) is fulfilled.

References

[1] B.D.Annin, G.P.Cherepanov, Elastic-plastic problem, Novosibirsk, Nauka, 1983 (in Russian).

[2] S.G.Lehnitsky, Theory of elasticity of an anisotropic body, Moscow, Nauka, 1977 (in Russian).

[3] S.I.Senashov, A.M.Vinogradov, *Proc. Edinburg Math. Soc.*, **31**(1988), no. 3, 415–439. DOI: 10.1017/S0013091500006817

[4] P.P.Kiryakov, S.I.Senashov, A.N.Yakhno, Application of symmetries and conservation laws to solving differential equations, Novosibirsk, Ros. Acad. nauk. Sib. otd., 2001 (in Russian).

[5] S.I.Senashov, O.V.Gomonova, A.N.Yakhno, Mathematical problems of two-dimensional equations of ideal plasticity, *Krasnoyarsk, Izd. SibGAU,* 2012 (in Russian).

Анизотропная антиплоская упругопластическая задача

Сергей И. Сенашов
Ирина Л. Савостьянова
Сибирский государственный университет науки и технологий им. М. Ф. Решетнева
Красноярск, Российская Федерация
Ольга Н. Черепанова
Сибирский федеральный университет
Красноярск, Российская Федерация

Аннотация. В работе решена анизотропная антиплоская упругопластическая задача о напряженном состоянии в теле, ослабленном отверстием, ограниченном кусочно-гладким контуром. В статье приведены законы сохранения, которые позволили свести вычисления компонент тензора напряжений к криволинейному интегралу по контуру отверстия. Законы сохранения дали возможность найти границу между упругой и пластической областями.

Ключевые слова: анизотропная упругопластическая задача, антиплоское напряженное состояние, законы сохранения.

DOI: 10.17516/1997-1397-2020-13-2-218-230
УДК 517.9

Asymptotic Analysis of Retrial Queueing System M/M/1 with Impatient Customers, Collisions and Unreliable Server

Elena Yu. Danilyuk*
Svetlana P. Moiseeva†
National Research Tomsk State University
Tomsk, Russian Federation
Janos Sztrik‡
University of Debrecen
Debrecen, Hungary

Received 29.11.2019, received in revised form 04.12.2019, accepted 20.01.2020

Abstract. The retrial queueing system of $M/M/1$ type with Poisson flow of arrivals, impatient customers, collisions and unreliable service device is considered in the paper. The novelty of our contribution is the inclusion of breakdowns and repairs of the service into our previous study to make the problem more realistic and hence more complicated. Retrial time of customers in the orbit, service time, impatience time of customers in the orbit, server lifetime (depending on whether it is idle or busy) and server recovery time are supposed to be exponentially distributed. An asymptotic analysis method is used to find the stationary distribution of the number of customers in the orbit. The heavy load of the system and long time patience of customers in the orbit are proposed as asymptotic conditions. Theorem about the Gaussian form of the asymptotic probability distribution of the number of customers in the orbit is formulated and proved. Numerical examples are given to show the accuracy and the area of feasibility of the proposed method.

Keywords: retrial queue, impatient customers, collisions, unreliable server, asymptotic analysis.

Citation: E.Yu.Danilyuk, S.P.Moiseeva, J.Sztrik, Asymptotic Analysis of Retrial Queueing System M/M/1 with Impatient Customers, Collisions and Unreliable Server, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 218–230. DOI: 10.17516/1997-1397-2020-13-2-218-230.

The ever increasing volume of information for designing communication systems in an optimal way requires new methods and approaches. More and more business processes involve big data transmission under limited capacities of devices. Therefore, developing of appropriate mathematical models of modern telecommunication systems and modifying of existing ones are very important. Queueing systems with repeated calls, or retrial queueing systems are suitable models for telecommunication systems. They are characterized by the feature that an arriving customer finding the server busy does not join a queue and does not leave the system immediately, but goes to some virtual place (orbit), and then it tries to get service again after some random time. A comprehensive description and comparison of classical queueing systems of retrial queues can be found in books by J. Artalejo and A. Gómez-Corral [1], J. Artalejo and G. Falin [2], G. Falin and J. Templeton [3], just to mentions some of them.

*daniluc.elena.yu@gmail.com https://orcid.org/0000-0002-7016-492X
†smoiseeva@mail.ru https://orcid.org/0000-0001-9285-1555
‡sztrik.janos@inf.unideb.hu https://orcid.org/0000-0002-5303-818X

The present paper generalizes the results obtained in [4,5]. We find the asymptotic stationary distribution of the number of calls in the orbit for the system under consideration. Collisions in the model usually arise in the analysis of communication networks when another message is transmitted during the transmission of a previous message. Such messages collide. They are considered distorted and both go into the orbit from where they ask the device for servicing again after a random delay time, see for example [4–8].

Impatience of calls in the orbit is understood as the case when a customer in the orbit can leave the orbit after a random time without service [4, 5, 7, 9–11]. But there is another way to specify impatience, for example, non-persistence, balking and reneging are used [8, 12–15]. Balking and reneging are fundamental concepts in queuing introduced by Anker, Gafarian [16], Haight [17] and Bareer [18]. They state that an arriving customer shows the least interest in joining a system which is already crowded. This behaviour is referred to as balking. Balking was applied to retention of reneged customers [19–23]. A comprehensive review of queueing systems with impatient customers can be found in [24].

In practice some components of the systems are subject to random breakdowns. Then it is very important to study reliability of retrial queues with server breakdowns and repairs because of the limited ability of repairs and heavy influence of the breakdowns on the performance of the system. Retrial queues with an unreliable server were studied , see for example [5, 14, 25, 26] and references therein.

More references on important papers devoted to the research of retrial queueing systems of various types (with impatient customers, collisions, and unreliable server) are given in our previous papers [4, 5, 7, 9–11, 25, 26].

The novelty of our contribution is the inclusion of breakdowns and repairs of the service into our previously developed models to make the problem more realistic and hence more complicated. In the present paper we continue to use an asymptotic analysis method developed at the Tomsk State University that is widely applied for the study of RQ -systems. This method makes it possible to obtain analytical result for different types of queueing systems and networks under specific asymptotic conditions.

The structure of the present work is as follows. Mathematical model of the novel retrial queueing system discussed in the paper and the problem statement are presented in Sect. 1. In Sect. 2 the detailed description of the model and the system of Kolmogorov equations for the stationary state probabilities are given. Sect. 3 consists of the solution of the problem by the asymptotic analysis method. Theorem on stationary probability distribution of the number in the orbit for retrial queueing system of M/M/1 type with impatient calls in the orbit, collisions and unreliable server under a long delay of calls in orbit and long time patience of calls in the orbit condition is formulated and proved in this section. Sect. 4 deals with some numerical examples that prove the theoretical results and illustrate the applicability of the proposed approach. Sect. 5 concludes the paper.

1. Description of the mathematical model

We consider a single server RQ-system with Poisson arrival process with parameter λ for the primary calls. A customer that finds the server idle takes it for service for an exponentially distributed random time with parameter μ. If the server is busy an arriving customer (either from the source or from the orbit) enters into a "collision" and both go into orbit. In the orbit each customer, called secondary calls, independently of others waits for a random time.

The waiting time is exponentially distributed with parameter σ. If the server is busy again the request tries to occupy the device to obtain servicing as soon as possible. If the server is idle the secondary customer occupies it for service for an exponentially distributed random time with parameter μ, that is, no difference between the service of primary and secondary calls.

We assume that server is unreliable, that is, the lifetime is supposed to be exponentially distributed with rate γ_0 if the server is idle and with parameter γ_1 if it is busy. When the server breaks down it is immediately sent for repair and the recovery time is assumed to be exponentially distributed with rate γ_2. When the server is down the primary sources continue generation of customers and send them to the server. Similarly, customers may retry from the orbit to the server but all arriving customers immediately go into the orbit. Furthermore, in this unreliable model we suppose that interrupted request goes to the orbit immediately and it's next service is independent of the interrupted one. All random variables involved in the model construction are assumed to be independent of each other.

Moreover, a customer in the orbit leaves the system without service after a random time which has an exponential distribution with rate α, demonstrating the "impatience" property. Fig. 1 shows the model of the RQ-system $M/M/1$ with impatient customers, collisions and unreliable server.

Our aims is to find the stationary distribution of the number of customers in the orbit for the described system.

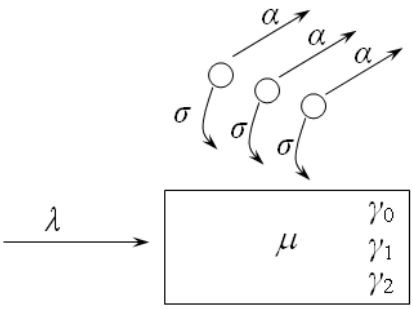

Fig. 1. Retrial queue M/M/1 with impatient customers in the orbit, collisions and unreliable server

2. System of the Kolmogorov differential equations

Let us consider the Markov process $\{k(t), i(t)\}$ of changing the states of the RQ-system under consideration, where $i(t)$ is the number of customers in the orbit at time t, $i(t) = 0, 1, 2, \ldots$ and $k(t)$ defines the state of the server at time t and takes one of three values:

$$k\,(t) = \begin{cases} 0, & \text{if the device is idle,} \\ 1, & \text{if the device is busy,} \\ 2, & \text{if the device is down (under repair).} \end{cases}$$

The joint probability that device is in state k at time t and i customers are in orbit is denoted by $P\{k(t) = k, i(t) = i\} = P(k, i, t)$.

To obtain the probability distribution $P\{k(t) = k, i(t) = i\} = P(k, i, t)$ for the states of the

considered RQ-system we derive a system of the Kolmogorov differential equations

$$\begin{cases} \dfrac{\partial P(0,i,t)}{\partial t} = -(\lambda + i\sigma + i\alpha + \gamma_0) P(0,i,t) + (i+1)\alpha P(0,i+1,t) + \\ \qquad + (i-1)\sigma P(1,i-1,t) + \mu P(1,i,t) + \lambda P(1,i-2,t) + \gamma_2 P(2,i,t), \\ \dfrac{\partial P(1,i,t)}{\partial t} = -(\lambda + \mu + i\sigma + i\alpha + \gamma_1) P(1,i,t) + \lambda P(0,i,t) + \\ \qquad + (i+1)\sigma P(0,i+1,t) + (i+1)\alpha P(1,i+1,t), \\ \dfrac{\partial P(2,i,t)}{\partial t} = -(\lambda + i\alpha + \gamma_2) P(2,i,t) + \lambda P(2,i-1,t) + \\ \qquad + (i+1)\alpha P(2,i+1,t) + \gamma_0 P(0,i,t) + \gamma_1 P(1,i-1,t), \end{cases} \tag{1}$$

$i = 0, 1, 2, \ldots$.

Since customers are "impatient" the considered system has a stationary distribution for any values of λ and μ. Let $\lim\limits_{t\to\infty} P(k,i,t) = P(k,i) = P_k(i)$, $k = \{0;1;2\}$. Then we can write system (1) in the following form

$$\begin{cases} -(\lambda + i\sigma + i\alpha + \gamma_0) P_0(i) + (i+1)\alpha P_0(i+1) + (i-1)\sigma P_1(i-1) + \mu P_1(i) + \\ \qquad + \lambda P_1(i-2) + \gamma_2 P_2(i) = 0, \\ -(\lambda + \mu + i\sigma + i\alpha + \gamma_1) P_1(i) + \lambda P_0(i) + (i+1)\sigma P_0(i+1) + (i+1)\alpha P_1(i+1) = 0, \\ -(\lambda + i\alpha + \gamma_2) P_2(i) + \lambda P_2(i-1) + (i+1)\alpha P_2(i+1) + \gamma_0 P_0(i) + \gamma_1 P_1(i-1) = 0, \\ \sum\limits_{i=0}^{\infty} (P_0(i) + P_1(i) + P_2(i)) = 1, \end{cases} \tag{2}$$

$i = 0, 1, 2, \ldots$.

System (2) is a system of difference equations of infinite dimension with variable coefficients. Such system is very difficult to solve. Therefore, to get solution we propose two approaches: asymptotic and numerical ones.

The numerical algorithm for finding the stationary probabilities is based on the reduction of dimension of system (2). To do that we represent system (2) for $i = 0, 1, 2, \ldots, N$ as

$$PS = B, \tag{3}$$

where the row vector P of dimension $3(N+1)$ is the desired stationary probability distribution of the number of customers in the orbit for each state of the device $k = \{0, 1, 2\}$

$$P = (P(0)\ P(1)\ P(2))$$

and $P(0)$, $P(1)$, $P(2)$ are row vectors with elements $P(0,i)$, $P(1,i)$, $P(2,i)$, $i = 0, 1, 2, \ldots, N$, respectively. Matrix S of dimension $3(N+1) \times (3(N+1)+1)$ is represented in the block form as

$$S = \begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \end{pmatrix},$$

where $S_{11} = ||s_{ij}^{11}||_1^{N+1}$, $S_{12} = ||s_{ij}^{12}||_1^{N+1}$, $S_{13} = ||s_{ij}^{13}||_1^{N+1}$, $S_{21} = ||s_{ij}^{21}||_1^{N+1}$, $S_{22} = ||s_{ij}^{22}||_1^{N+1}$, $S_{23} = ||s_{ij}^{23}||_1^{N+1}$, $S_{31} = ||s_{ij}^{31}||_1^{N+1}$, $S_{32} = ||s_{ij}^{32}||_1^{N+1}$, $S_{33} = ||s_{ij}^{33}||_1^{N+1}$ are sparse matrices with

non-zero elements defined as

$$s_{ii}^{11} = -\lambda - (i-1)(\sigma+\alpha) - \gamma_0, \quad s_{i+1,i}^{11} = i\alpha,$$
$$s_{ii}^{21} = \mu, \quad s_{i,i+1}^{21} = (i-1)\sigma, \quad s_{i,i+2}^{21} = \lambda,$$
$$s_{ii}^{31} = \gamma_2,$$
$$s_{ii}^{12} = \lambda, \quad s_{i,i-1}^{12} = i\sigma,$$
$$s_{ii}^{22} = -(\lambda + \mu + \gamma_1 + (i-1)(\sigma+\alpha)), \quad s_{i+1,i}^{22} = i\alpha,$$
$$s_{ii}^{13} = \gamma_0,$$
$$s_{i,i+1}^{23} = \gamma_1,$$
$$s_{ii}^{33} = -(\lambda + \gamma_2 + (i-1)\alpha), \quad s_{i+1,i}^{33} = i\alpha, \quad s_{i,i+1}^{33} = \lambda.$$

Blocks S_{14}, S_{24}, S_{34} are unit vector columns of dimension $(N+1)$, row vector $B = ||b_i||$ of dimension $3(N+1)+1$ is a row of free coefficients with elements $b_n = 0$ $(n = 0, 1, 2, \ldots, N-1)$, $b_N = 1$.

We solve (3) with the use of the Mathcad software package. We choose N to be so large that probabilities $P(0, N)$, $P(1, N)$, $P(2, N)$ are equal to the machine zero.

The numerical algorithm provides satisfactory accuracy but it has a drawback due to the limited computational capacity of the computer. Therefore, analytical methods are needed to calculate the probability distribution of the number of customers in orbit for the considered RQ-system with impatient customers, collisions and unreliable server. They allow us to find a distribution for a system of any dimension. Thus the alternative to the numerical method is the method of asymptotic analysis.

3. Asymptotic analysis

To find the solution of system of equations (2) we propose another approach by using the method of asymptotic analysis under the assumption that there is a long delay between customers from the orbit and high "patience" customers, i.e., when $\sigma \to 0$, $\alpha \to 0$. We summarize the results of our study in the next Theorem 3.1.

Theorem 3.1. *The stationary distribution of the number of customers in the orbit in the RQ-system M/M/1 with impatient customers, collisions and unreliable server under the above assumptions and conditions $\alpha = q\sigma$, $q > 0$ is the asymptotically normal distribution with mean G_1/σ and variance G_2/σ. Here*

$$G_1 = \frac{\lambda - \mu R_1}{q}, \tag{4}$$

$$G_2 = \frac{G_1\,(q + R_0) + \gamma_1 f_0 + \lambda R_1}{R_0 - R_1 + q}, \tag{5}$$

$$f_1 = \frac{(2\lambda + (1-q)\,G_1)\,R_1 - (1+q)\,G_1 R_0}{\gamma_0 + \gamma_1 + 2\gamma_2}. \tag{6}$$

R_1 is the probability that the server is busy in the stationary regime of system operation. It is determined by equation

$$(\gamma_0 + \gamma_1 + 2\gamma_2)\,\mu R_1^2 - cR_1 + \lambda\,(1+q)\,\gamma_2 = 0, \quad R_1 \in [0; 1]\,, \tag{7}$$

$$c = (1+q)\,\lambda\,(\gamma_0+\gamma_1+2\gamma_2) + q\,(\gamma_0+\gamma_2)\,(\mu+\gamma_1) + \mu\gamma_2.$$

R_0 is the probability that the server is idle in the stationary regime of system operation. It is determined by equation

$$R_0 = \frac{\gamma_2 - (\gamma_1+\gamma_2)\,R_1}{\gamma_0+\gamma_2}, \quad R_0 \in [0;1]\,. \tag{8}$$

Proof. The method of asymptotic analysis in queueing theory is the method of study of the equations to determine some characteristics of an queueing system under some limit (asymptotic) condition which is specific for any model and problem under consideration.

We introduce the partial characteristic functions as follows

$$H_k(u) = \sum_{i=0}^{\infty} e^{jui} P_k(i), \quad H_k(0) = \sum_{i=0}^{\infty} P_k(i) \triangleq R_k, \tag{9}$$

where $j = \sqrt{-1}$, $k = \{0,1,2\}$, and R_k are stationary state probabilities of the process $k(t)$. It is obvious that $H(u) = H_0(u) + H_1(u) + H_2(u)$.

Using (9) and $H_k'(u) = \dfrac{dH_k(u)}{du} = j\sum\limits_{i=0}^{\infty} ie^{jui}P_k(i)$, $k = \{0,1,2\}$, we can write system (2) as follows

$$\begin{cases} -(\lambda+\gamma_0)\,H_0(u) + j\,(\sigma+\alpha)\,H_0'(u) + \mu H_1(u) - j\alpha e^{-ju}H_0'(u) - j\sigma e^{ju}H_1'(u) + \\ \qquad + \lambda e^{2ju}H_1(u) + \gamma_2 H_2(u) = 0, \\ -(\lambda+\mu+\gamma_1)\,H_1(u) + j\,(\sigma+\alpha)\,H_1'(u) + \lambda H_0(u) - j\alpha e^{-ju}H_1'(u) - j\sigma e^{-ju}H_0'(u) = 0, \\ -(\lambda+\gamma_2)\,H_2(u) + j\alpha H_2'(u) + \lambda e^{ju}H_2(u) - j\alpha e^{-ju}H_2'(u) + \gamma_0 H_0(u) + \gamma_1 e^{ju}H_1(u) = 0. \end{cases} \tag{10}$$

Adding the first and the second equations by the third one of (10) we get the system below

$$\begin{cases} -(\lambda+\gamma_0)\,H_0(u) + \left(\mu+\lambda e^{2ju}\right)H_1(u) + \gamma_2 H_2(u) + j\left(\sigma+\alpha-\alpha e^{-ju}\right)H_0'(u) - \\ \qquad - j\sigma e^{ju}H_1'(u) = 0, \\ \lambda H_0(u) - (\lambda+\mu+\gamma_1)\,H_1(u) - j\sigma e^{-ju}H_0'(u) + j\left(\sigma+\alpha-\alpha e^{-ju}\right)H_1'(u) = 0, \\ \gamma_0 H_0(u) + \gamma_1 e^{ju}H_1(u) - \left(\lambda+\gamma_2-\lambda e^{ju}\right)H_2(u) + j\alpha\left(1-e^{-ju}\right)H_2'(u) = 0, \\ \left[\lambda\left(e^{ju}+1\right)+\gamma_1\right]H_1(u) + \lambda H_2(u) + je^{-ju}\,(\sigma+\alpha)\,H_0'(u) + j\left(\alpha e^{-ju}-\sigma\right)H_1'(u) + \\ \qquad + j\alpha e^{-ju}H_2'(u) = 0. \end{cases} \tag{11}$$

System (11) is the basic system for further analysis of retrial queueing system of $M/M/1$ type with impatient customers in the orbit, collisions and unreliable server under a long delay of customers in orbit ($\sigma \to 0$) and long time patience of customers in the orbit ($\alpha \to 0$) conditions.

The proof of Theorem 3.1 is carried out in two stages.

Stage 1. Finding the first-order asymptotic.

Let us make the substitutions $\sigma = \varepsilon$, $\alpha = q\varepsilon$, $u = \varepsilon w$, $H_k(u) = = F_k(w,\varepsilon)$, $k = \{0,1,2\}$ in basic system (11), where $\varepsilon \to 0$.

Since $H_k'(u) = \frac{1}{\varepsilon}\frac{\partial F_k(w,\varepsilon)}{\partial w}$, $k = \{0,1,2\}$ systems of equations (11) can be written as

$$\begin{cases} -(\lambda+\gamma_0)F_0(w,\varepsilon) + \left(\mu+\lambda e^{2jw\varepsilon}\right)F_1(w,\varepsilon) + \gamma_2 F_2(w,\varepsilon) + \\ \quad + j\left(1+q-qe^{-jw\varepsilon}\right)\dfrac{\partial F_0(w,\varepsilon)}{\partial w} - je^{jw\varepsilon}\dfrac{\partial F_1(w,\varepsilon)}{\partial w} = 0, \\ \lambda F_0(w,\varepsilon) - (\lambda+\mu+\gamma_1)F_1(w,\varepsilon) - je^{-jw\varepsilon}\dfrac{\partial F_0(w,\varepsilon)}{\partial w} + j\left(1+q-qe^{-jw\varepsilon}\right)\dfrac{\partial F_1(w,\varepsilon)}{\partial w} = 0, \\ \gamma_0 F_0(w,\varepsilon) + \gamma_1 e^{jw\varepsilon}F_1(w,\varepsilon) - \left(\lambda+\gamma_2-\lambda e^{jw\varepsilon}\right)F_2(w,\varepsilon) + jq\left(1-e^{-jw\varepsilon}\right)\dfrac{\partial F_2(w,\varepsilon)}{\partial w} = 0, \\ \left[\lambda\left(e^{jw\varepsilon}+1\right)+\gamma_1\right]F_1(w,\varepsilon) + \lambda F_2(w,\varepsilon) + je^{-jw\varepsilon}(1+q)\dfrac{\partial F_0(w,\varepsilon)}{\partial w} + \\ \quad + j\left(qe^{-jw\varepsilon}-1\right)\dfrac{\partial F_1(w,\varepsilon)}{\partial w} + jqe^{-jw\varepsilon}\dfrac{\partial F_2(w,\varepsilon)}{\partial w} = 0. \end{cases} \tag{12}$$

The transformation of equations (12) under $\varepsilon \to 0$ with $F_k(w) = \lim\limits_{\varepsilon\to 0} F_k(w,\varepsilon)$, $F_k'(w) = \dfrac{dF_k(w)}{dw}$, $k = \{0,1,2\}$ leads to

$$\begin{cases} -(\lambda+\gamma_0)F_0(w) + (\mu+\lambda)F_1(w) + \gamma_2 F_2(w) + jF_0'(w) - jF_1'(w) = 0, \\ \lambda F_0(w) - (\lambda+\mu+\gamma_1)F_1(w) - jF_0'(w) + jF_1'(w) = 0, \\ \gamma_0 F_0(w) + \gamma_1 F_1(w) - \gamma_2 F_2(w) = 0, \\ \lambda\left[F_0(w)+F_1(w)+F_2(w)\right] - \mu F_1(w) + jq\left[F_0'(w)+F_1'(w)+F_2'(w)\right] = 0. \end{cases} \tag{13}$$

We seek solution of equation (13) $F_k(w)$, $k = \{0,1,2\}$ in the form

$$F_k(w) = R_k\Phi(w), \quad k = \{0,1,2\}, \tag{14}$$

where R_0, R_1, R_2 are defined in (9), $R_0 + R_1 + R_2 = 1$, $R_k = H_k(0) = F_k(0)$, $k = \{0,1,2\}$, and $\Phi(w)$ is an unknown function.

Substituting (14) into (13), we obtain the system of differential equations with respect to function $\Phi(w)$

$$\begin{cases} \left[-(\lambda+\gamma_0)R_0 + (\mu+\lambda)R_1 + \gamma_2 R_2\right]\Phi(w) = j(R_1-R_0)\dfrac{d\Phi(w)}{dw}, \\ \left[\lambda R_0 - (\lambda+\mu+\gamma_1)R_1\right]\Phi(w) = j(R_1-R_0)\dfrac{d\Phi(w)}{dw}, \\ \gamma_0 R_0\Phi(w) + \gamma_1 R_1\Phi(w) - \gamma_2 R_2\Phi(w) = 0, \\ \left[\lambda(R_0+R_1+R_2) - \mu R_1\right]\Phi(w) = -jq\left[R_0+R_1+R_2\right]\dfrac{d\Phi(w)}{dw}. \end{cases} \tag{15}$$

According to equations (15), we can find

$$\Phi(w) = \exp\left\{G_1 jw\right\}, \tag{16}$$

where G_1 is given in (4). It follows from the forth equation (15).

It is obvious that solution of system (15) exists when the following equalities are satisfied

$$\begin{cases} (\lambda+\gamma_0)R_0 - (\mu+\lambda)R_1 - \gamma_2 R_2 = \lambda R_0 - (\lambda+\mu+\gamma_1)R_1, \\ \gamma_0 R_0 + \gamma_1 R_1 - \gamma_2 R_2 = 0, \\ R_0 + R_1 + R_2 = 1. \end{cases} \tag{17}$$

Expressions for $R_0, R_1, R_2 \in [0;1]$ can be obtained from system of equations (17). They are

$$R_0 = \frac{\gamma_2 - (\gamma_1 + \gamma_2) R_1}{\gamma_0 + \gamma_1}, \quad R_2 = \frac{\gamma_0 - (\gamma_0 - \gamma_1) R_1}{\gamma_0 + \gamma_2}, \tag{18}$$

and R_1 is the root of equation

$$(\gamma_0 + \gamma_1 + 2\gamma_2)\,\mu R_1^2 - cR_1 + \lambda\,(1+q)\,\gamma_2 = 0, \tag{19}$$

$$c = (1+q)\,\lambda\,(\gamma_0 + \gamma_1 + 2\gamma_2) + q\,(\gamma_0 + \gamma_2)\,(\mu + \gamma_1) + \mu\gamma_2.$$

Equation (19) has at least one root $R_1 \in [0;1]$, and the proof of existence of R_1 is similar to that in [4].

Using (14), (16) and $\varepsilon = \sigma$ we can write the expression for the partial characteristic functions as follows

$$H_k(u) = F_k(w,\varepsilon) = F_k(w) + o(\varepsilon) \approx F_k(w) = F_k\left(\frac{u}{\varepsilon}\right) = R_k \exp\left\{\frac{G_1}{\sigma} ju\right\}, \tag{20}$$

$k = \{0,1,2\}$, and $R_0, R_1, R_2 \in [0;1]$ are defined in (18), (19).

Using (20) and $R_0 + R_1 + R_2 = 1$ the pre-limit characteristic function $H(u) = H_0(u) + H_1(u) + H_2(u)$ under the assumption of a long delay of customers in the orbit and their high "patience" can be approximated by function $h_1(u)$

$$h_1(u) = \exp\left\{\frac{G_1}{\sigma} ju\right\}, \tag{21}$$

which is the first-order asymptotic characteristic function (or the first-order asymptotic).

Stage 2. Finding the second-order asymptotic.

Taking into account (21), we assume

$$H_k(u) = \exp\left\{\frac{G_1}{\sigma} ju\right\} H_k^{(2)}(u), \quad k = \{0,1,2\}, \tag{22}$$

in basic system of equations (11). Then systems of equations (11) can be rewritten as follows

$$\begin{cases} -\left[\lambda + \gamma_0 + \dfrac{G_1}{\sigma}\left(\sigma + \alpha - \alpha e^{-ju}\right)\right] H_0^{(2)}(u) + \left(\mu + \lambda e^{2ju} + G_1 e^{ju}\right) H_1^{(2)}(u) + \\ \qquad + \gamma_2 H_2^{(2)}(u) + j\left(\sigma + \alpha - \alpha e^{-ju}\right) \dfrac{dH_0^{(2)}(u)}{du} - j\sigma e^{ju} \dfrac{dH_1^{(2)}(u)}{du} = 0, \\ \left(\lambda + G_1 e^{-ju}\right) H_0^{(2)}(u) - \left[\lambda + \mu + \gamma_1 + \dfrac{G_1}{\sigma}\left(\sigma + \alpha - \alpha e^{-ju}\right)\right] H_1^{(2)}(u) - \\ \qquad - j\sigma e^{-ju} \dfrac{dH_0^{(2)}(u)}{du} + j\left(\sigma + \alpha - \alpha e^{-ju}\right) \dfrac{dH_1^{(2)}(u)}{du} = 0, \\ \gamma_0 H_0^{(2)}(u) + \gamma_1 e^{ju} H_1^{(2)}(u) - \left[\lambda + \gamma_2 - \lambda e^{ju} + \dfrac{G_1\alpha}{\sigma}\left(1 - e^{-ju}\right)\right] H_2^{(2)}(u) + \\ \qquad + j\alpha\left(1 - e^{-ju}\right) \dfrac{dH_2^{(2)}(u)}{du} = 0, \\ -e^{-ju}\dfrac{G_1}{\sigma}\left(\sigma+\alpha\right) H_0^{(2)}(u) + \left[\lambda\left(e^{ju}+1\right) + \gamma_1 - \dfrac{G_1}{\sigma}\left(\alpha e^{-ju} - \sigma\right)\right] H_1^{(2)}(u) + \left(\lambda - \dfrac{G_1\alpha}{\sigma} e^{-ju}\right) \times \\ \quad \times H_2^{(2)}(u) + je^{-ju}\left(\sigma+\alpha\right) \dfrac{dH_0^{(2)}(u)}{du} + j\left(\alpha e^{-ju} - \sigma\right) \dfrac{dH_1^{(2)}(u)}{du} + j\alpha e^{-ju} \dfrac{dH_2^{(2)}(u)}{du} = 0. \end{cases} \tag{23}$$

Let $\sigma = \varepsilon^2, \alpha = q\varepsilon^2$, $u = \varepsilon w$, $H_k^{(2)}(u) = F_k^{(2)}(w,\varepsilon)$, $k = \{0,1,2\}$, where $\varepsilon \to 0$. Then system (23) after some transformations becomes

$$\begin{cases} -\left[\lambda+\gamma_0+G_1\left(1+q-qe^{-jw\varepsilon}\right)\right]F_0^{(2)}(w,\varepsilon)+\left(\mu+\lambda e^{2jw\varepsilon}+G_1e^{jw\varepsilon}\right)F_1^{(2)}(w,\varepsilon)+ \\ \quad +\gamma_2F_2^{(2)}(w,\varepsilon)+j\varepsilon\left(1+q-qe^{-jw\varepsilon}\right)\dfrac{\partial F_0^{(2)}(w,\varepsilon)}{\partial w}-j\varepsilon e^{jw\varepsilon}\dfrac{\partial F_1^{(2)}(w,\varepsilon)}{\partial w}=0, \\ \left(\lambda+G_1e^{-jw\varepsilon}\right)F_0^{(2)}(w,\varepsilon)-\left[\lambda+\mu+\gamma_1+G_1\left(1+q-qe^{-jw\varepsilon}\right)\right]F_1^{(2)}(w,\varepsilon)- \\ \quad -j\varepsilon e^{-jw\varepsilon}\dfrac{\partial F_0^{(2)}(w,\varepsilon)}{\partial w}+j\varepsilon\left(1+q-qe^{-jw\varepsilon}\right)\dfrac{\partial F_1^{(2)}(w,\varepsilon)}{\partial w}=0, \\ \gamma_0F_0^{(2)}(w,\varepsilon)+\gamma_1e^{jw\varepsilon}F_1^{(2)}(w,\varepsilon)-\left[\lambda+\gamma_2-\lambda e^{jw\varepsilon}+G_1q\left(1-e^{-jw\varepsilon}\right)\right]F_2^{(2)}(w,\varepsilon)+ \\ \quad +jq\varepsilon\left(1-e^{-jw\varepsilon}\right)\dfrac{\partial F_2^{(2)}(w,\varepsilon)}{\partial w}=0, \\ -e^{-jw\varepsilon}G_1(1+q)F_0^{(2)}(w,\varepsilon)+\left[\lambda\left(e^{jw\varepsilon}+1\right)+\gamma_1-G_1\left(qe^{-jw\varepsilon}-1\right)\right]F_1^{(2)}(w,\varepsilon)+ \\ \quad +\left(\lambda-G_1qe^{-jw\varepsilon}\right)F_2^{(2)}(w,\varepsilon)+j\varepsilon e^{-jw\varepsilon}(1+q)\dfrac{\partial F_0^{(2)}(w,\varepsilon)}{\partial w}+ \\ \quad +j\varepsilon\left(qe^{-jw\varepsilon}-1\right)\dfrac{\partial F_1^{(2)}(w,\varepsilon)}{\partial w}+jq\varepsilon e^{-jw\varepsilon}\dfrac{\partial F_2^{(2)}(w,\varepsilon)}{\partial w}=0. \end{cases} \tag{24}$$

When $\varepsilon \to 0$ in (24) and $\lim\limits_{\varepsilon\to 0} F_k^{(2)}(w,\varepsilon) = F_k^{(2)}(w)$, $k = \{0,1,2\}$ we obtain

$$\begin{cases} -(\lambda+\gamma_0+G_1)F_0^{(2)}(w)+(\mu+\lambda+G_1)F_1^{(2)}(w)+\gamma_2F_2^{(2)}(w)=0, \\ (\lambda+G_1)F_0^{(2)}(w)-(\lambda+\mu+\gamma_1+G_1)F_1^{(2)}(w)=0, \\ \gamma_0F_0^{(2)}(w)+\gamma_1F_1^{(2)}(w)-\gamma_2F_2^{(2)}(w)=0, \\ -G_1(1+q)F_0^{(2)}(w)+[2\lambda+\gamma_1-G_1(q-1)]F_1^{(2)}(w)+(\lambda-G_1q)F_2^{(2)}(w)=0. \end{cases} \tag{25}$$

The solution of systems of equations (24) has the following form

$$\begin{cases} F_k^{(2)}(w,\varepsilon)=(R_k+jw\varepsilon f_k)\,\Phi^{(2)}(w)+o\left(\varepsilon^2\right), \quad k=\{0,1,2\}, \\ R_0+R_1+R_2=1, \end{cases} \tag{26}$$

where R_0, R_1, R_2 are defined above, f_0, f_1, f_2 are constants and function $\Phi^{(2)}(w)$ is to be determined.

Substituting (26) into (24) and taking into account (25), with the proviso that $\varepsilon \to 0$ one can write the system as

$$\begin{cases} [(\lambda+\gamma_0+G_1)f_0+G_1qR_0-(\mu+\lambda+G_1)f_1-(2\lambda+G_1)R_1-\gamma_2f_2]\,w\Phi^{(2)}(w)= \\ =(R_0-R_1)\dfrac{d\Phi^{(2)}(w)}{dw}, \\ [(\lambda+G_1)f_0-G_1R_0-(\lambda+\mu+\gamma_1+G_1)f_1-G_1qR_1]\,w\Phi^{(2)}(w)=(R_0-R_1)\dfrac{d\Phi^{(2)}(w)}{dw}, \\ [\gamma_0f_0+\gamma_1f_1+\gamma_1R_1-\gamma_2f_2-(G_1q-\lambda)R_2]\,w\Phi^{(2)}(w)=0, \\ [(2\lambda+\gamma_1+G_1-G_1q)f_1+(\lambda+G_1q)R_1-(\lambda-G_1q)f_2+G_1qR_2-G_1(1+q)f_0+ \\ \quad +G_1(1+q)R_0]\,w\Phi^{(2)}(w)=[(1-q)R_1-(1+q)R_0-qR_2]\dfrac{d\Phi^{(2)}(w)}{dw}, \end{cases} \tag{27}$$

where R_0, R_1 and G_1 are defined in (18), (19) and (16), respectively.

The solution of system (27) has the form

$$\Phi^{(2)}(w) = \exp\left\{G_2 \frac{(jw)^2}{2}\right\}, \tag{28}$$

where G_2 is defined in (5).

Using the same transformation as for the first-order asymptotic and additional conditions $f_2 - f_1 - f_0 = 0$ and $f_1 - f_0 = 0$, we finally obtain expressions for the solution of system (27)

$$\begin{cases} (\lambda + \gamma_0 + G_1) f_0 + G_1 q R_0 - (\mu + \lambda + G_1) f_1 - (2\lambda + G_1) R_1 - \gamma_2 f_2 = \\ \qquad = (\lambda + G_1) f_0 - G_1 R_0 - (\lambda + \mu + \gamma_1 + G_1) f_1 - G_1 q R_1, \\ \gamma_0 f_0 + \gamma_1 f_1 + \gamma_1 R_1 - \gamma_2 f_2 - (G_1 q - \lambda) R_2 = 0, \\ 2f_1 + f_2 = 0, \\ f_1 - f_0 = 0. \end{cases} \tag{29}$$

Making the reverse substitutions we obtain

$$H_k^{(2)}(u) = F_k^{(2)}(w, \varepsilon) = (R_k + jw\varepsilon f_k) \exp\left\{G_2 \frac{(jw)^2}{2}\right\} + o\left(\varepsilon^2\right) \approx R_k \exp\left\{\frac{G_2}{\sigma} \frac{(ju)^2}{2}\right\}. \tag{30}$$

Then using (30), expressions (22) can be written as

$$H_k(u) = \exp\left\{\frac{G_1}{\sigma} ju\right\} H_k^{(2)}(u) \approx R_k \exp\left\{\frac{G_1}{\sigma} ju + \frac{G_2}{\sigma} \frac{(ju)^2}{2}\right\}, \quad k = \{0, 1, 2\}. \tag{31}$$

Taking into account (31), characteristic function $H(u) = H_0(u) + H_1(u) + H_2(u)$. Assuming that customers in the orbit have long delays and the "patience" is high, we can see that distribution is the Gaussian one. Hence

$$h_2(u) = \exp\left\{\frac{G_1}{\sigma} ju + \frac{G_2}{\sigma} \frac{(ju)^2}{2}\right\}. \tag{32}$$

Theorem 3.1 is proved. □

4. Numerical results

In this section we give some comments to Theorem 3.1 and several numerical examples are considered.

We construct asymptotic distributions of the probabilities of the number of customers in the orbit with parameters $\mu = 1$, $\gamma_0 = 0.1$, $\gamma_1 = 0.2$, $\gamma_2 = 1$, $\alpha = 2\sigma$ for various values of the delay parameter σ and parameter λ. Then these distributions are compared with pre-limit (numerical) distributions obtained by the matrix method.

Fig. 2 shows one of samples for $\lambda = 0.7$ and $\sigma = 0.01$ (left picture) and $\sigma = 0.001$ (right picture).

As a measure of proximity of two distributions the Kolmogorov distance

$$\Delta = \max_{0 \leqslant i \leqslant N} \left| \sum_{k=0}^{i} P_{matrix}(i) - \sum_{k=0}^{i} P_{asimpt}(i) \right|$$

is used, where $P_{matrix}(i)$ is the probability distribution of the number of customers in the orbit obtained by the matrix method, $P_{asimpt}(i)$ is the asymptotic probability distribution of the number of customers in the orbit.

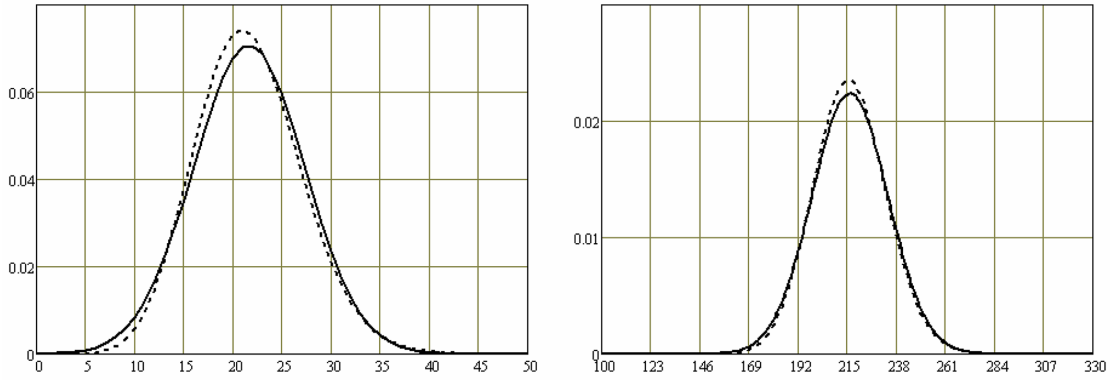

Fig. 2. Asymptotic (dashed line) and pre-limit(numerical) (solid line) probability distributions of the number of calls in the orbit

Table 1. Values of the Kolmogorov distance

λ/μ	Kolmogorov distance Δ				
	$\sigma = 0.1$	$\sigma = 0.05$	$\sigma = 0.01$	$\sigma = 0.005$	$\sigma = 0.001$
0.5	0.161	0.101	0.023	0.016	0.009
0.7	0.117	0.066	0.020	0.016	0.013
0.9	0.092	0.048	0.021	0.020	0.018
1.1	0.075	0.039	0.024	0.023	0.022
1.5	0.055	0.035	0.030	0.029	0.029
2.0	0.046	0.037	0.035	0.035	0.035

Conclusion

Retrial queueing system of $M/M/1$ type with impatient customers in the orbit, collisions and unreliable server is considered in the paper. It is proved that stationary probability distribution of the number of customers in the orbit can be approximated by the Gaussian distribution under conditions of a long delay and a long patience time of the customers in the orbit. The accuracy of the approximation was compared with numerical results obtained with the use of the matrix analytic method.

The study was funded by Russian Foundation for Basic Research and Tomsk region (project no. 19-41-703002).

References

[1] J.R.Artalejo, A.Gomez-Corral, Retrial Queueing Systems: A Computational Approach, Springer, 2008.

[2] J.R.Artalejo, G.I.Falin, Standard and retrial queueing systems: A comparative analysis, *Revista Matematica Complutense*, **15**(2002), 101–129.

[3] G.I.Falin, J.G.C.Templeton, Retrial queues, London, New York: Chapman & Hall, 1997.

[4] E.Yu.Danilyuk, E.A.Fedorova, S.P.Moiseeva, *Automation and Remote Control*, **79**(2018), no. 12, 2136–2146. DOI: 10.1134/S0005117918120044

[5] A.Nazarov, J.Sztrik, A.Kvach, T.Berczes, *Annals of Operations Research*, **277**(2019), no. 2, 213–229. DOI: 10.1007/s10479-018-2894-z

[6] L.Lakaour, D.Aïssani, K.Adel-Aissanou, K.Barkaoui, *Methodology and Computing in Applied Probability*, **21**(2019), no. 4, 1395–1406. DOI: 10.1007/s11009-018-9680-x

[7] E.Danilyuk, S.Moiseeva, A.Nazarov, In: Dudin A., Nazarov A., Moiseev A. (eds). Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2019, Communications in Computer and Information Science, Vol. 1109, 2019, 230–242. DOI: 10.1007/978-3-030-33388-1_19

[8] J.Kim, Retrial queueing system with collision and impatience, *Communications of the Korean Mathematical Society*, *Communications of the Korean Mathematical Society*, **4**(2010), 647–653.

[9] O.Vygovskaya, E.Danilyuk, S.Moiseeva, In: Dudin A., Nazarov A., Moiseev A. (eds). Queueing Theory and Applications. ITMM 2018, WRQ 2018. Communications in Computer and Information Science, Vol. 912, 2018, 387–399. DOI: 10.1007/978-3-319-97595-5_30

[10] E.Danilyuk, O.Vygovskaya, S.Moiseeva, S.Rozhkova, In: Vishnevskiy V., Kozyrev D. (eds). DCCN 2018. Communications in Computer and Information Science, Vol. 919, 2018, 493–504. DOI: 10.1007/978-3-319-99447-5_42

[11] E.Fedorova, E.Danilyuk, A.Nazarov, A.Melikov, In: Phung-Duc T., Kasahara S., Wittevrongel S. (eds) Queueing Theory and Network Applications. QTNA 2019. Lecture Notes in Computer Science, Vol. 11688, 2019, 3–15. DOI: 10.1007/978-3-030-27181-7_1

[12] M.P.D'Arienzo, A.N.Dudin, S.A.Dudin, R.Manzo, *Journal of Ambient Intelligence and Humanized Computing*, (2019), 1–9. DOI: 10.1007/s12652-019-01318-x

[13] B.Kim, J.Kim, Extension of the loss probability formula to an overloaded queue with impatient customers, *Statistics & Probability Letters*, **134**(2018), 54–62. DOI: 10.1016/j.spl.2017.10.007

[14] A.Aissani, F.Lounis, D.Hamadouche, S.Taleb, *American Journal of Mathematical and Management Sciences*, **38**(2019), no. 2, 125–150. DOI: 10.1080/01966324.2018.1486763

[15] A.N.Dudin, Operations Research Perspectives, Operations Research Perspectives, *Operations Research*, **5**(2018), 245–255.

[16] C.J.Ancker, A.V.Gafarian, Some Queuing Problems with Balking and Reneging. I, *Operations Research*, **11**(1963), no. 1, 88–100.

[17] F.A.Haight, Queueing with Balking, *Biometrika*, **44**(1957), no. 3/4, 360–369.

[18] D.Y.Barrer, Queuing with Impatient Customers and Indifferent Clerks, *Operations Research*, **5**(1957), no. 5, 644–649.

[19] R.Kumar, S.K.Sharma, *International Journal of Mathematics in Operational Research*, **5**(2013), no. 6, 709–720. DOI: 10.1504/ijmor.2013.057488

[20] R.Kumar, B.K.Som, Economic analysis of M/M/c/N queue with retention of impatient customers, *Advance Modeling and optimization*, **16**(2014), no. 2, 339–353.

[21] B.K.Som, *Advances in Analytics and Applications*, (2018), 261–272. DOI: 10.1007/978-981-13-1208-3_20

[22] A.Santhakumaran, B.Thangaraj, A Single Server Queue with Impatient and Feedback Customers, *Information and Management Sciences*, **11**(2000), no. 3, 71–79.

[23] B.K.Som, Cost-profit analysis of stochastic heterogeneous queue with reverse balking, feedback and retention of impatient customers, *Reliability: Theory and Applications*, **14**(2019), no. 1(52), 87–101.

[24] K.Wang, N.Li, Z.Jiang, Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, 2010, 82–87. DOI: 10.1109/SOLI.2010.5551611

[25] A.Nazarov, J.Sztrik, A.Kvach, In: Dudin,A., Nazarov, A. (eds.). Queueing Theory and Applications. ITMM 2017, CCIS, Vol. 800, 2017, 97–110. DOI: 10.1007/978-3-319-68069-9

[26] T.Berczes, J.Sztrik, A.Toth, A.Nazarov, In: Dudin,A., Nazarov, A. (eds.). Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2017, CCIS, Vol. 800, 2017, 248–258. DOI: 10.1007/978-3-319-68069-9

Асимптотический анализ системы массового обслуживания с повторными вызовами M/M/1 с нетерпеливыми заявками, конфликтами и ненадежным прибором

Елена Ю. Данилюк
Светлана П. Моисеева
Национальный исследовательский Томский государственный университет
Томск, Российская Федерация
Янош Стрик
Университет Дебрецена
Дебрецен, Венгрия

Аннотация. В настоящей статье мы рассматриваем систему массового обслуживания с повторными вызовами (RQ-систему) типа M/M/1 с пуассоновским потоком поступающих в систему заявок и одним сервером, обслуживание которым имеет экспоненциальное распределение. Классическая модель RQ-системы усложнена наличием конфликтов заявок в системе, "нетерпеливых" заявок на орбите, а также "ненадежным" прибором, который выходит из строя и ремонтируется в функционирующей системе массового обслуживания. Время, через которое заявки с орбиты вновь обращаются к обслуживающему прибору; время, через которое заявки с орбиты покидают систему, время, в течение которого сервер находится в рабочем состоянии (в зависимости от того, занят прибор обслуживанием заявки или нет, а также время, в течение которого длится ремонт вышедшего из строя сервера, распределены экспоненциально. Мы используем метод асимптотического анализа для решения задачи нахождения распределения вероятностей числа заявок на орбите. В качестве асимптотического условия предлагается условие высокой загрузки системы и долгой "терпеливости" заявок на орбите. Формулируется и доказывается теорема об асимптотически гауссовском распределении вероятностей числа заявок на орбите. Приводятся численные результаты, демонстрирующие область применения полученных теоретических выводов.

Ключевые слова: RQ-система, нетерпеливые заявки, конфликты, ненадежный прибор, асимптотический анализ.

DOI: 10.17516/1997-1397-2020-13-2-231-241
УДК 519.716

E-closed Sets of Hyperfunctions on Two-Element Set

Vladimir I. Panteleyev*
Leonid V. Riabets†
Irkutsk State University
Irkutsk, Russian Federation

Received 09.11.2019, received in revised form 06.01.2020, accepted 13.02.2020

Abstract. Hyperfunctions are functions that are defined on a finite set and return all non-empty subsets of the considered set as their values. This paper deals with the classification of hyperfunctions on a two-element set. We consider the composition and the closure operator with the equality predicate branching (E-operator). E-closed sets of hyperfunctions are sets that are obtained using the operations of adding dummy variables, identifying variables, composition, and E-operator. It is shown that the considered classification leads to a finite set of closed classes. The paper presents all 78 E-closed classes of hyperfunctions, among which there are 28 pairs of dual classes and 22 self-dual classes. The inclusion diagram of the E-closed classes is constructed, and for each class its generating system is obtained.

Keywords: closure, equality predicate, hyperfunction, closed set, composition.

Citation: V.I.Panteleyev, L.V.Riabets, E-closed Sets of Hyperfunctions on Two-Element Set, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 231–241. DOI: 10.17516/1997-1397-2020-13-2-231-241.

Systems with generalization of k-valued logic functions have been studied for a long time along with classical functional systems over a set of k-valued functions ($k \geqslant 2$). Such systems are based on partial functions, multifunctions, hyperfunctions. These functions are defined on a finite set A and take values in the set of subsets of A. Usually such systems are closed with respect to the composition operator (see [1–7]).

The composition operator leads to a countable or continuous classification; therefore, closure operators that generate finite classifications of functions are of interest. Such operators, in particular, include the parametric and positive closure operators [8], the operator with the equality predicate branching (E-operator) [9]. An investigation of E-operator on the set of Boolean functions, partial Boolean functions and on the set of functions of k-valued logic can be found in [9–11]. All E-closed classes for the set of partial Boolean functions were obtained in [12]. The complete structure of closed classes for parametric and positive closure operators for hyperfunctions on two-element set was obtained in [13, 14]. The completeness criterion for the E-operator on the set of hyperfunctions of rank two was proved in [15].

The aim of this paper is to describe all E-closed classes of hyperfunctions on a two-element set.

Introduction

Let $E_2 = \{0, 1\}$ and $\mathcal{P}(E_2)$ be the power set of E_2. An n-ary hyperfunction f on E_2 is a mapping

*vl.panteleyev@gmail.com https://orcid.org/0000-0003-4766-486X
†l.riabets@gmail.com https://orcid.org/0000-0003-4047-9573

$$f : E_2^n \to \mathcal{P}(E_2) \setminus \{\varnothing\}.$$

We will write P_2^- for $\mathcal{P}(E_2) \setminus \{\varnothing\}$. Let $H_{2,n} = (P_2^-)^{E_2^n}$ be the set of all n-ary hyperfunctions on A, $n \geqslant 1$, and $H_2 = \bigcup_n H_{2,n}$ be the set of all finitary hyperfunctions on E_2.

An i-th n-ary projection (a selector hyperfunction) on E_2, $1 \leqslant i \leqslant n$, is the n-ary hyperfunction $e_i^n \in H_{2,n}$ defined by $e_i^n(x_1, \ldots, x_n) = \{x_i\}$.

In what follows, we will not distinguish between a set of one element and an element of this set. For the set E_2, we will use the notation "$-$" (dash).

An n-variable hyperfunction f will be represented as a vector $(\tau_{\tilde{0}}, \ldots, \tau_{\tilde{1}})$, where $\tilde{0}, \ldots, \tilde{1}$ are binary representations of numbers $0, \ldots, 2^n - 1$ and $\tau_{\tilde{\sigma}}$ equals to $f(\tilde{\sigma})$. Such vectors have the form $(f(0)\, f(1))$ for unary hyperfunctions and $(f(0,0)\, f(0,1)\, f(1,0)\, f(1,1))$ for binary hyperfunctions.

Let $f \in H_{2,n}$ and $f_1, \ldots, f_n \in H_{2,m}$ for positive integers n and m. The composition of hyperfunctions f and $f_1, \ldots, f_n$ is an n-ary hyperfunction $f(f_1, \ldots, f_n)$ defined by

$$f(f_1, \ldots, f_n)(\alpha_1, \ldots, \alpha_m) = \bigcup_{\beta_i \in f_i(\alpha_1, \ldots, \alpha_m)} f(\beta_1, \ldots, \beta_n),$$

where $(\alpha_1, \ldots, \alpha_m) \in E_2^m$.

We say that a hyperfunction $g(x_1, \ldots, x_n)$ is obtained from the functions $f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n)$ using the operator with the equality predicate branching (E-operator) if for some $i, j \in \{1, \ldots, n\}$ the following relation holds:

$$g(x_1, \ldots, x_n) = \begin{cases} f_1(x_1, \ldots, x_n) & \text{if } x_i = x_j, \\ f_2(x_1, \ldots, x_n) & \text{otherwise.} \end{cases}$$

The set of all hyperfunctions H_2 that can be obtained from the set $Q \subseteq H_2$ using the operations of adding dummy variables, identifying variables, composition and equality predicate branching is called the E-closure of set Q.

A set of hyperfunctions that coincides with its closure is called an E-closed class. We say that the set $R \subseteq Q$ E-generates an E-closed class Q if E-closure of the set R coincides with the class Q. Therefore R is an E-complete in Q. Following [10] the E-closure of R is denoted by $[R]_E$. By $Q(n)$ we denote a set of all n-variable hyperfunctions from Q.

A hyperfunction g is called dual to a hyperfunction f if $g(x_1, \ldots, x_n) = \overline{f}(\overline{x}_1, \ldots, \overline{x}_n)$. A class that includes all hyperfunctions dual to hyperfunctions of class Q, is called dual to the class Q and denoted $\overline{Q}$. The class Q will be called self-dual if $Q = \overline{\overline{Q}}$.

The original hyperfunction and the hyperfunctions obtained from it by identifying variables or adding dummy variables will be denoted by the same symbols, if this does not cause confusion.

1. The extended operator with the equality predicate branching

The operator with the equality predicate branching allows us to reduce the E-closure of H_2 to the E-closure of sets of 2-variable hyperfunctions.

Proposition 1.1. *Any E-closed $Q \subseteq H_2$ is E-generated by the set of all its hyperfunctions depending on at most two variables.*

Proof. The proof is similar to the proof of the corresponding statement in [10].

Proposition 1.1 shows that there are finite number of E-closed classes in H_2. Moreover, for any two E-closed classes Q_1 and Q_2, $Q_1 \subseteq Q_2$ is equivalent to $Q_1(2) \subseteq Q_2(2)$. Thus, the problem of describing all E-closed classes in H_2 is reduced to the construction of all E-closed sets of 2-variable hyperfunctions. Before presenting an algorithm for constructing such sets let us pay attention to the following fact (the similar fact is mentioned in [12]).

Consider hyperfunctions $f_1, \ldots, f_k$ depending on no more than n variables. If we try to obtain n-variable hyperfunction f using $f_1, \ldots, f_k$, we may need to use intermediate hyperfunctions that depend on more than n variables.

As an example, let $f_1(x_1, x_2) = (-101)$, $f_2(x_1, x_2) = (-011)$, and $f_3(x_1, x_2) = (-1\ -\ 1)$. Applying composition operator and E-operator to f_1, f_2, f_3 without using hyperfunctions of a larger number of variables allows us to obtain only f_1, f_2, or f_3.

Let $g(x_1, x_2, x_3)$ be a hyperfunction such that

$$g(x_1, x_2, x_3) = \begin{cases} f_1(x_1, x_2) & \text{if } x_1 = x_3, \\ f_2(x_1, x_2) & \text{otherwise.} \end{cases}$$

Now we can obtain $f(x_1, x_2) = (-0 - 1)$:

$$f(x_1, x_2) = g(x_1, x_2, f_3(x_2, x_2)).$$

Thus, to work with hyperfunctions of no more than two variables, it is necessary to make more precise the definitions of the used operators.

Let f, g_1, g_2, h_1, and h_2 be 2-variable hyperfunctions. We say that the hyperfunction f is obtained from the functions g_1, g_2, h_1, h_2 using an extended operator with the equality predicate branching (Ex-operator) if for any binary set $(\alpha_1, \alpha_2) \in E_2^2$ the following relation holds:

- if $h_1(\alpha_1, \alpha_2), h_2(\alpha_1, \alpha_2) \in E_2$, then

$$f(\alpha_1, \alpha_2) = \begin{cases} g_1(\alpha_1, \alpha_2) & \text{if } h_1(\alpha_1, \alpha_2) = h_2(\alpha_1, \alpha_2), \\ g_2(\alpha_1, \alpha_2) & \text{otherwise;} \end{cases}$$

- if $h_1(\alpha_1, \alpha_2) = -$ or $h_2(\alpha_1, \alpha_2) = -$, then

$$f(\alpha_1, \alpha_2) = g_1(\alpha_1, \alpha_2) \cup g_2(\alpha_1, \alpha_2).$$

For brevity, we use the notation:

$$f(x_1, x_2) = Ex\big(g_1(x_1, x_2), g_2(x_1, x_2), h_1(x_1, x_2), h_2(x_1, x_2)\big).$$

Further, we will consider the composition operator (restricted composition) only in the following form:

$$g_1(h_1(x_1, x_2), h_2(x_1, x_2)).$$

For the above-defined operators hyperfunctions h_1 and h_2 can be selector hyperfunctions. The closure of the set Q obtained with respect to the extended operator with the equality predicate branching, restricted composition, operation of adding dummy variables, and identifying variables will be denoted by $[Q]_{Ex}$.

2. Ex-closed classes of H_2

The definition of Ex-closure allows us to formulate an algorithm for constructing Ex-closed classes of hyperfunctions.

The algorithm constructs Ex-generated sets of hyperfunctions. Each 2-variable hyperfunction will be associated with its number from 0 to 80. At each iteration, a sequence of restricted compositions and a sequence of extended operations with the equality predicate branching are applied.

The algorithm builds one-element and two-element Ex-generated sets separately. Computer calculations showed that there are no three-element Ex-generated sets. We describe the steps of the algorithm in the form of pseudo-code.

```
function get_hyperfunction_classes() {
vars
   F: collection<function>;
   Q: collection<class>;
   A: class;
for each f in H_2 do {
   A = new class;
   A ← f;
   while has_new(A) do {
      A ← composition(A);
      A ← Ex(A);
   }
   if is_new(Q, A) then{
      Q ← A;
      F ← f;
   }
}
while has_new(Q) do {
   for each B in Q do {
      for each f in F do {
         B ← f;
         while has_new(B) do {
            B ← composition(B);
            B ← Ex(B);
         }
         if is_new(Q, B) then{
            Q ← B;
         }
      }
   }
}
return Q;
}
```

The algorithm has been implemented in Java. It was found that there are precisely 78 Ex-closed classes of H_2. Among them, there are 56 classes that are divided into pairs of pairwise

dual classes and 22 classes are self-dual.

Now we list the known E-closed classes of H_2 obtained in [15]:
$T_0^{0-} = \{f(x_1, \ldots, x_n) \mid f(0, \ldots, 0) \in \{0, -\}\}$;
$T_1^{1-} = \{f(x_1, \ldots, x_n) \mid f(1, \ldots, 1) \in \{1, -\}\}$;
$S^- = \{f(x_1, \ldots, x_n) \mid \big(f(\alpha_1, \ldots, \alpha_n), f(\overline{\alpha}_1, \ldots, \overline{\alpha}_n)\big) \notin \{(0,0),(1,1)\}, (\alpha_1, \ldots, \alpha_n) \in E_2^n\}$;
$O_2 = \{f(x_1, \ldots, x_n) \mid f(\alpha_1, \ldots, \alpha_n) \in \{0,1\}, (\alpha_1, \ldots, \alpha_n) \in E_2^n\}$.

According to the main theorem in [15] these classes are E-precomplete in H_2.

In [9] it was shown that the well-known classes of Boolean functions T_0, T_1, S, T_{01}, S_{01}, C_0, and C_1 are E-closed.

The self-dual classes $U_1, \ldots, U_{16}$, H_2, S^-, O_2, S, S_{01}, T_{01} and non-dual classes $V_1, \ldots, V_{25}$, T_0^{0-}, T_0, C_0 (one from each pair) are presented in the form of an inclusion diagram in Fig 1.

Additional information on Ex-closed classes is presented in Tab. 1. The first column shows the name of Ex-closed class, the second column shows the number of 2-variable hyperfunctions in the class, and the third column shows the class generating system. One representative class from the pair of pairwise dual classes is presented in Tab. 1.

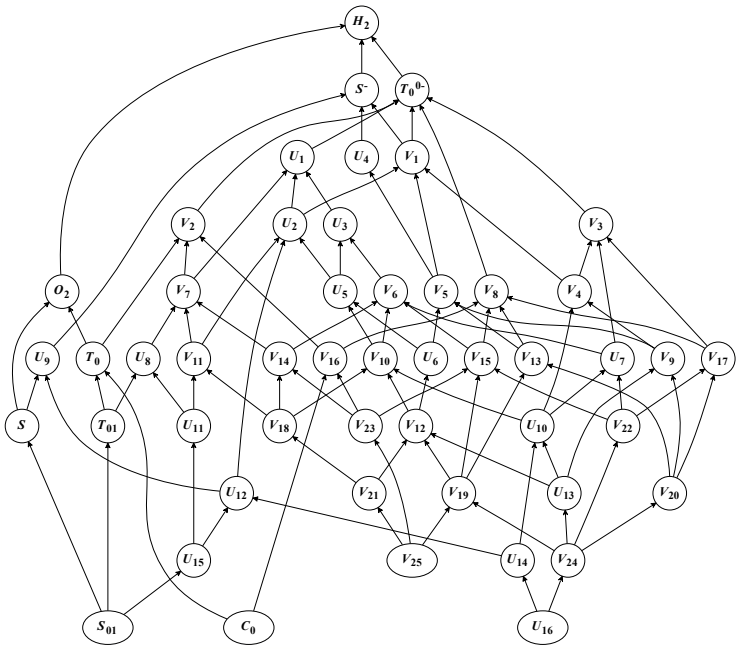

Fig. 1. The diagram of inclusions for Ex-closed classes of H_2

3. E-closed classes of H_2

The result obtained in the previous section allows us to formulate an upper bound for E-closed classes of hyperfunctions.

Theorem 3.1. *For any set $Q \subseteq H_2$, it follows that $[Q]_{Ex} \subseteq [Q]_E$.*

Proof. We show that each hyperfunction obtained using extended operators with the equality predicate branching and restricted composition can be represented by a formula using the E-closure operators.

Table 1. The generating sets for Ex-closed classes

1	**2**	**3**
H_2	81	$(0000), (1-00)$
T_0^{0-}	54	$(0000), (-100)$
S^-	49	$(1-00)$
U_1	36	$(0001), (-00-)$
V_1	35	$(-100), (0101)$
U_2	28	$(0101), (--0-)$
U_3	27	$(-001), (000-)$
V_2	27	$(0-10)$
V_3	27	(-110)
U_4	25	$(--00), (1-0-)$
U_5	21	$(-101), (010-)$
V_4	21	(-100)
V_5	20	$(--00), (0-1-)$
V_6	18	$(000-), (-10-)$
V_7	18	$(0001), (000-)$
V_8	16	$(0000), (-000)$
O_2	16	(1000)
U_6	15	$(--01), (0-0-)$
V_9	15	$(--10)$
V_{10}	14	$(010-), (-10-)$
V_{11}	14	$(0101), (010-)$
V_{12}	10	$(0-0-), (--1-)$
U_7	9	$(-00-), (-10-)$
U_8	9	$(0001), (0-01)$
V_{13}	9	$(--00), (0-0-)$
V_{14}	9	$(011-)$
U_9	9	$(1--0)$
V_{15}	8	$(000-), (-00-)$
V_{16}	8	$(0-00)$
T_0	8	(0100)
V_{17}	8	(-000)
U_{10}	7	$(-10-), (--0-)$
U_{11}	7	$(0-01)$
V_{18}	7	$(010-)$
V_{19}	6	$(0-0-), (--0-)$
U_{12}	6	$(0101), (-10-)$
V_{20}	6	$(--00)$
U_{13}	5	$(--0-), (--1-)$
V_{21}	5	$(0-1-)$
V_{22}	4	$(-00-)$
V_{23}	4	$(000-)$
T_{01}	4	(0001)
S	4	(1100)
V_{24}	3	$(--0-)$
U_{14}	3	$(-10-)$
U_{15}	3	$(0--1)$
V_{25}	3	$(0-0-)$
S_{01}	2	(0101)
U_{16}	1	$(----)$
C_0	1	(0000)

Consider

$$f(x,y,z,t) = \begin{cases} g_1(x,y) & \text{if } z = t, \\ g_2(x,y) & \text{otherwise.} \end{cases}$$

Let

$$U(x,y) = f(x,y,h_1(x,y),h_2(x,y)).$$

Now we obtain its possible values on some binary set (α_1,α_2) for various hyperfunctions h_1 and h_2. Let $h_1(\alpha_1,\alpha_2) = \tau_1$ and $h_2(\alpha_1,\alpha_2) = \tau_2$. Consider all possible values for τ_1 and τ_2.

- Let $\tau_1 \in E_2$ and $\tau_2 \in E_2$. Then

$$U(\alpha_1,\alpha_2) = f(\alpha_1,\alpha_2,\tau_1,\tau_2) = \begin{cases} g_1(\alpha_1,\alpha_2) & \text{if } \tau_1 = \tau_2, \\ g_2(\alpha_1,\alpha_2) & \text{otherwise.} \end{cases}$$

- Let $\tau_1 = -$ and3 $\tau_2 \in E_2$. Substitute these values

$$U(\alpha_1,\alpha_2) = f(\alpha_1,\alpha_2,-,\tau_2) = f(\alpha_1,\alpha_2,0,\tau_2) \cup f(\alpha_1,\alpha_2,1,\tau_2) = g_1(\alpha_1,\alpha_2) \cup g_2(\alpha_1,\alpha_2).$$

The case $\tau_1 \in E_2$ and $\tau_2 = -$ is similar to the previous one.

- Let $\tau_1 = -$ and $\tau_2 = -$. Then

$$U(\alpha_1, \alpha_2) = f(\alpha_1, \alpha_2, -, -) = \\ = f(\alpha_1, \alpha_2, 0, 0) \cup f(\alpha_1, \alpha_2, 0, 1) \cup f(\alpha_1, \alpha_2, 1, 0) \cup f(\alpha_1, \alpha_2, 1, 1) = \\ = g_1(\alpha_1, \alpha_2) \cup g_2(\alpha_1, \alpha_2).$$

Therefore the values of $U(\alpha_1, \alpha_2)$ coincide with the values of an extended operator with the equality predicate branching based on the functions g_1, g_2, h_1, and h_2. This proves that $[Q]_E \supseteq [Q]_{Ex}$. □

Corollary 1. *The number of E-closed classes is at most 78.*

To obtain a lower bound for the number of E-closed classes, we consider the following sets of hyperfunctions:

$K_1 = T_0^{0-} = \{f(x_1, \dots, x_n) \mid f(0, \dots, 0) \in \{0, -\}\}$;
$K_2 = \{f(x_1, \dots, x_n) \mid f(0, \dots, 0) = 0\}$;
$K_3 = \{f(x_1, \dots, x_n) \mid f(0, \dots, 0) = -\}$;
$K_4 = T_1^{1-} = \{f(x_1, \dots, x_n) \mid f(1, \dots, 1) \in \{1, -\}\}$;
$K_5 = \{f(x_1, \dots, x_n) \mid f(1, \dots, 1) = 1\}$;
$K_6 = \{f(x_1, \dots, x_n) \mid f(1, \dots, 1) = -\}$;
$K_7 = O_2$ is a set of Boolean functions;
$K_8 = \{f(x_1, \dots, x_n) \mid f(\alpha_1, \dots, \alpha_n) \in \{0, -\}, (\alpha_1, \dots, \alpha_n) \in E_2^n\}$;
$K_9 = \{f(x_1, \dots, x_n) \mid f(\alpha_1, \dots, \alpha_n) \in \{1, -\}, (\alpha_1, \dots, \alpha_n) \in E_2^n\}$;
$K_{10} = S^-$ is a set of self-dual hyperfunctions;

K_{11} is a set of hyperfunctions such that for $R = \{(01), (10), (--)\}$, for any n, and for any $(\alpha_{11}, \alpha_{12}), \dots, (\alpha_{n1}, \alpha_{n2}) \in R$, it follows that $\big(f(\alpha_{11}, \dots, \alpha_{n1}), f(\alpha_{12}, \dots, \alpha_{n2})\big) \in R$;

K_{12} is a set of hyperfunctions that take values $(0-), (-0), (1-), (-1)$, or $(--)$ on any pair of opposite binary sets;

$K_{13} = \{f \mid f \in T_0^{0-} \cap T_1^{1-}$ and $- \in \{f(0, \dots, 0), f(1, \dots, 1)\}\}$.

Proposition 3.1. *The sets $K_1, \dots, K_{13}$ are pairwise coinciding.*

It is evident that sets K_1–K_{10} are E-closed classes.

Lemma 3.1. *The set K_{11} is an E-closed class.*

Proof. Let a hyperfunction f be obtained by a composition of hyperfunctions $g, g_1, \dots, g_m \in K_{11}$. Suppose that $f \notin K_{11}$. Thus, on the sets of R, it takes the values $(0-)$, (-0), $(1-)$, (-1), (00), or (11). At the same time, R contains only three sets. It can be assumed that f depends on three variables.

Now consider the case for the pair $(0-)$. Let $f(01-) = 0$ and $f(10-) = -$. From the second equality it follows that there is a binary set (10α) such that $f(10\alpha) = 0$ or $f(10\alpha) = -$. By definition, if $f(01-) = 0$, then $f(01\overline{\alpha}) = 0$. Thus on binary sets $f(01\overline{\alpha}) = 0$ and $f(10\alpha) = 0$ or $f(01\overline{\alpha}) = 0$ and $f(10\alpha) = -$. Note also that $(00) \notin R$ and $(0-) \notin R$. At the same time $(\overline{\alpha}\alpha) \in R$.

On the other hand, consider the composition $g(g_1(x_1, x_2, x_3), \dots, g_m(x_1, x_2, x_3))$ on sets $(01\overline{\alpha})$ and (10α). Since $g, g_1, \dots, g_m \in K_{11}$, we have $\big(g_1(01\overline{\alpha})g_1(10\alpha)\big), \dots, \big(g_1(01\overline{\alpha})g_1(10\alpha)\big) \in R$. It follows that

$$\big(g(g_1(01\overline{\alpha}), \dots, g_m(01\overline{\alpha})), g(g_1(10\alpha), \dots, g_m(10\alpha))\big) \in R.$$

We get a contradiction. The remaining pairs (-0), $(1-)$, (-1), (00), and (11) are verified similarly.

Consider the operator with the equality predicate branching. Let

$$f(x_1,\dots,x_n)=\begin{cases} g_1(x_1,\dots,x_n) & \text{if } x_i=x_j,\\ g_2(x_1,\dots,x_n) & \text{otherwise,}\end{cases}$$

where $g_1, g_2 \in K_{11}$.

Consider the value of $f(x_1,\dots,x_n)$ on sets of R. By definition, if $\alpha_s \in \{0,1,-\}$, $s=\overline{1,n}$, then

$$f(\alpha_1,\dots,\alpha_n)=\bigcup_{\beta_s\in\alpha_s} f(\beta_1,\dots,\beta_n)=\bigcup_{\beta_i=\beta_j} g_1(\beta_1,\dots,\beta_n)\bigcup\bigcup_{\beta_i\neq\beta_j} g_2(\beta_1,\dots,\beta_n).$$

Since $\overline{\beta}_s\in\overline{\alpha}_s \Leftrightarrow \beta_s\in\alpha_s$ and $\beta_s=\beta_t \Leftrightarrow \overline{\beta}_s=\overline{\beta}_t$, $t\in\{1,\dots,n\}$, we have

$$f(\overline{\alpha}_1,\dots,\overline{\alpha}_n)=\bigcup_{\gamma_s\in\overline{\alpha}_s} f(\gamma_1,\dots,\gamma_n)=\bigcup_{\substack{\beta_i=\beta_j\\ \beta_s\in\alpha_s}} g_1(\overline{\beta}_1,\dots,\overline{\beta}_n)\bigcup\bigcup_{\substack{\beta_i\neq\beta_j\\ \beta_s\in\alpha_s}} g_2(\overline{\beta}_1,\dots,\overline{\beta}_n).$$

It is easily shown that the set R is closed with respect to the operation of joining sets. In other words, if $(\alpha_1,\alpha_2)\in R$ and $(\beta_1,\beta_2)\in R$, then $(\alpha_1\cup\beta_1,\alpha_2\cup\beta_2)\in R$. This completes the proof. □

Lemma 3.2. *The set K_{12} is an E-closed class.*

Proof. The validity of the statement for the operator with the equality predicate branching is obvious. We can use the assumption of contradiction to prove that K_{12} is closed with respect to composition. □

It can be shown in the usual way that K_{13} is an E-closed class.

Theorem 3.2. *Ex-closed classes are E-closed.*

Proof. Let us prove that the described above Ex-closed classes are different with respect to the E-closure. For each Ex-closed class K, we construct the vector $v_K=(\gamma_K^1,\dots,\gamma_K^{13})$, which indicates that the class K is a subset of K_1-K_{13} :

$$\gamma_K^i=\begin{cases}1 \text{ if } K\subseteq K_i,\\ 0 \text{ otherwise.}\end{cases}$$

Clearly, if $\gamma_K^j=1$ and $\gamma_{K'}^j=0$, then $[K]_E\neq[K']_E$.

For the convenience of comparison, we divide the set of all vectors v_K into four groups with respect to hyperfunctions belonging to the precomplete classes T_0^{0-} (E-closed class K_1) and T_1^{1-} (E-closed class K_4). By $K_1\overline{K}_4$ we denote the set of hyperfunctions belonging to K_1 and not belonging to K_4. The other sets we denote as K_1K_4, $\overline{K}_1K_4$, $\overline{K}_1\overline{K}_4$. Note also that Ex-closed classes belonging to different groups are different with respect to the E-closure. In the tables below, we replace the character 0 with an empty cell in each of v_K.

Group $\overline{\mathbf{K}}_\mathbf{1}\overline{\mathbf{K}}_\mathbf{4}$. The classes of this group are distinguished by the sets K_7, K_{10}, K_{11}, K_{12} (see Tab. 2).

Group $\overline{\mathbf{K}}_\mathbf{1}\mathbf{K}_\mathbf{4}$ and group $\mathbf{K}_\mathbf{1}\overline{\mathbf{K}}_\mathbf{4}$. The sets in these classes are dual; therefore, if there exists a hyperfunction $f\in K$ from one group, then there exists a hyperfunction $f^*\in K^*$ from another group. Thus, if the hyperfunctions of the first group are distinguished by the sets $M_1,\dots,M_t$, then dual functions of the second group are distinguished by the dual sets $M_1^*,\dots,M_t^*$.

The classes of $\mathbf{K}_\mathbf{1}\overline{\mathbf{K}}_\mathbf{4}$ are distinguished by the sets K_2, K_3, K_7, K_8, K_{10}, K_{12} (see Tab. 3).

Table 2. Classes of $\overline{K}_1\overline{K}_4$

N		B	K_7	K_{10}	K_{11}	K_{12}
1	S	(1100)	1	1	1	
2	U_9	$(1--0)$		1	1	
3	O_2	(1000)	1			

N		B	K_7	K_{10}	K_{11}	K_{12}
4	U_4	$(--00), (1-0-)$		1		1
5	S^-	$(1-00)$		1		
6	H_2	$(0000), (1-00)$				

Table 3. Classes of $K_1\overline{K}_4$

N		B	K_2	K_3	K_7	K_8	K_{10}	K_{12}
7	C_0	(0000)	1		1	1		
8	V_{20}	$(--00)$		1		1	1	1
9	V_{17}	(-000)		1		1		
10	V_{16}	$(0-00)$	1			1		
11	V_{13}	$(--00), (0-0-)$				1	1	1
12	V_8	$(0000), (-000)$				1		
13	T_0^{0-}	$(0000), (-100)$						

N		B	K_2	K_3	K_7	K_8	K_{10}	K_{12}
14	V_5	$(--00), (0-1-)$					1	1
15	V_4	(-100)		1			1	
16	V_3	(-110)		1				
17	V_2	$(0-10)$	1					
18	V_1	$(-100), (0101)$					1	
19	V_9	$(--10)$		1			1	1
20	T_0	(0100)	1		1			

The group $\overline{\mathbf{K_1}}\mathbf{K_4}$ contains 14 sets dual to the sets of $\mathbf{K_1}\overline{\mathbf{K}}_\mathbf{4}$: C_1, V_{20}^*, V_{17}^*, V_{16}^*, V_{13}^*, V_8^*, T_1^{1-}, V_5^*, V_4^*, V_3^*, V_2^*, V_1^*, V_9^*, T_1. These classes are distinguished by the sets K_5, K_6, K_7, K_9, K_{10}, K_{12}.

Group $\mathbf{K_1K_4}$. We divide $K_1 \cap K_4$ into two subsets with respect to belonging to the E-precomplete class of self-dual hyperfunctions S^- (E-closed class K_{10}). The sets are presented in Tabs. 4 and 5. The enumeration of sets is performed, taking into account 14 sets of $\overline{\mathbf{K_1}}\mathbf{K_4}$. This completes the proof. □

Table 4. Classes of $K_1 \cap K_4$ and not belonging to K_{10}

N		B	K_2	K_3	K_5	K_6	K_7	K_8	K_9	K_{13}
35	V_{15}	$(000-)$, $(-00-)$				1		1		1
36	V_{23}	$(000-)$	1			1		1		1
37	T_{01}	(0001)	1		1		1			
38	V_{14}	$(011-)$	1			1				1
39	U_8	(0001), $(0-01)$	1		1					
40	V_7^*	(0001), (-001)			1					
41	V_7	(0001), $(000-)$	1							
42	V_6	$(000-)$, $(-10-)$				1				1

N		B	K_2	K_3	K_5	K_6	K_7	K_8	K_9	K_{13}
43	V_{23}^*	(-111)		1	1				1	1
44	V_{22}	$(-00-)$		1		1		1		1
45	V_{22}^*	$(-11-)$		1		1			1	1
46	V_{15}^*	(-111), $(-11-)$		1					1	1
47	V_{14}^*	(-001)		1	1					1
48	U_7	$(-00-)$, $(-10-)$		1		1				1
49	V_6^*	(-001), $(-00-)$		1						1
50	U_1	(0001), $(-00-)$								
51	U_3	(-001), $(000-)$								1

Theorem 3.3. *The number of E-closed classes of H_2 is 78.*

Proof. It follows from Theorem 3.1 and Theorem 3.2. □

Table 5. Classes of $K_1 \cap K_4$ and belonging to K_{10}

N		B	K_2	K_3	K_5	K_6	K_7	K_8	K_9	K_{11}	K_{12}	K_{13}
52	U_{16}	$(----)$		1		1		1	1	1	1	1
53	S_{01}	(0101)	1		1		1			1		
54	V_{25}^*	$(--11)$		1	1				1		1	1
55	U_{15}	$(0--1)$	1		1					1		
56	U_{14}	$(-10-)$		1		1				1		1
57	V_{25}	$(0-0-)$	1			1		1			1	1
58	V_{24}	$(--0-)$		1		1		1			1	1
59	V_{24}^*	$(--1-)$		1		1			1		1	1
60	V_{21}^*	$(--01)$		1	1						1	1
61	V_{21}	$(0-1-)$	1			1					1	1
62	U_{13}	$(--0-), (--1-)$		1		1					1	1
63	U_{12}	$(0101), (-10-)$								1		
64	V_{19}^*	$(--11), (--1-)$		1					1		1	1
65	V_{19}	$(0-0-), (--0-)$				1		1			1	1
66	V_{18}^*	(-101)		1	1							1
67	U_{11}	$(0-01)$	1		1							
68	V_{18}	$(010-)$	1			1						1
69	U_{10}	$(-10-), (--0-)$		1		1						1
70	V_{12}^*	$(--01), (--0-)$		1							1	1
71	V_{12}	$(0-0-), (--1-)$				1					1	1
72	V_{11}^*	$(0101), (-101)$			1							
73	V_{11}	$(0101), (010-)$	1									
74	V_{10}^*	$(-101), (-10-)$		1								1
75	V_{10}	$(010-), (-10-)$				1						1
76	U_6	$(--01), (0-0-)$									1	1
77	U_5	$(-101), (010-)$										1
78	U_2	$(0101), (--0-)$										

References

[1] D.Lau, Function algebras on finite sets. A basic course on many-valued logic and clone theory, 2006.

[2] H.Machida, *Multiple-Valued Logic*, **8**(2002), 495–501. DOI: 10.1080/10236620215294

[3] H.Machida, J.Pantovic, On Maximal Hyperclones on $\{0,1\}$ — a new approach, Proceedings of 38th IEEE International Symposium on Multiple-Valued Logic, 2008, 32–37.

[4] B.A.Romov, Hyperclones on a Finite Set, *Multiple-Valued Logic*, **3**(1998), 285–300.

[5] V.I.Panteleyev, Completeness Criterion for Additional Defined Boolean Functions, *Vestnik Sam. Gos. Univ.*, **68**(2009), 60–79 (in Russian).

[6] S.V.Zamaratskaya, V.I.Panteleyev, Maximal Clones of Ultrafunctions of Rank 2, *The Bulletin of Irkutsk State University. Series Mathematics*, **15**(2016), 26–37 (in Russian).

[7] S.V.Zamaratskaya, V.I.Panteleyev, Classification and Types of Bases of All Ultrafunctions on Two-Element Set, *The Bulletin of Irkutsk State University. Series Mathematics*, **16**(2016), 58–70 (in Russian).

[8] S.S.Marchenkov, *Discrete Math. Appl.*, **9**(1999), no. 6, 563–581.
DOI: 10.1515/dma.1999.9.6.563

[9] S.S.Marchenkov, Closure Operators with Predicate Branching, *Vestnik Moskov. Univ. Ser. 1. Mat. Mekh.*, 2003, no. 6, 37–39 (in Russian).

[10] S.S.Marchenkov, The Closure Operator With the Equality Predicate Branching on the Set of Partial Boolean Functions, *Discrete Math. Appl.*, **18**(2008), no. 4, 381–389.

[11] S.S.Marchenkov, E-closed Operator in the Set of Partial Many-Valued Logic Functions, *Mathematical problems in cybernetics*, (2013), no. 19, 227–238 (in Russian).

[12] S.A.Matveev, Construction of All E-closed Classes of Partial Boolean Functions, *Mathematical problems in cybernetics*, (2013), no.18, 239–244 (in Russian).

[13] L.V.Ryabets, Parametric Closed Classes of Hyperfunctions on Two-Element Set, *The Bulletin of Irkutsk State University. Series Mathematics*, **17**(2016), 46–61 (in Russian).

[14] L.V.Ryabets, Parametric and Positive Closed Operators on the Set of Hyperfunctions on Two-Element Set, *Intelligent systems. Theory and applications*, **20**(2016), no. 3, 79–84 (in Russian).

[15] V.I.Panteleyev, L.V.Ryabets, The Closure Operator with the Equality Predicate Branching on the Set of Hyperfunctions on Two-Element Set, *The Bulletin of Irkutsk State University. Series Mathematics*, **10**(2014), 93–105 (in Russian).

E-замкнутые классы гиперфункций ранга 2

Владимир И. Пантелеев
Леонид В. Рябец
Иркутский государственный университет
Иркутск, Российская Федерация

Аннотация. Гиперфункции представляют собой функции, задаваемые на конечном множестве и возвращающие в качестве своих значений все непустые подмножества рассматриваемого множества. В работе рассматривается классификация гиперфункций, заданных на двухэлементном множестве, относительно оператора *E*-замыкания. *E*-замкнутыми множествами гиперфункций являются множества, замкнутые относительно суперпозиции, оператора замыкания с разветвлением по предикату равенства, отождествления переменных и добавления фиктивных переменных. Показано, что рассматриваемая классификация приводит к конечному множеству замкнутых классов. В работе описаны все 78 *E*-замкнутых классов гиперфункций, среди которых есть 28 пар двойственных классов и 22 самодвойственных класса. Построена диаграмма включений классов и для каждого класса указана его порождающая система.

Ключевые слова: замыкание, предикат равенства, гиперфункция, замкнутое множество, суперпозиция.

DOI: 10.17516/1997-1397-2020-13-2-242-252
УДК 519.622

First-Order Methods With Extended Stability Regions for Solving Electric Circuit Problems

Mikhail V. Rybkov*
Lyudmila V. Knaub†
Danil V. Khorov‡
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 16.01.2020, received in revised form 06.02.2020, accepted 25.03.2020

Abstract. Stability control of Runge-Kutta numerical schemes is studied to increase efficiency of integrating stiff problems. The implementation of the algorithm to determine coefficients of stability polynomials with the use of the GMP library is presented. Shape and size of the stability region of a method can be preassigned using proposed algorithm. Sets of first-order methods with extended stability domains are built. The results of electrical circuits simulation show the increase of the efficiency of the constructed first-order methods in comparison with methods of higher order.

Keywords: stiff problem, explicit methods, stability region, accuracy and stability control.

Citation: M.V.Rybkov, L.V.Knaub, D.V.Khorov, First-Order Methods With Extended Stability Regions for Solving Electric Circuit Problems, J. Sib. Fed. Univ. Math. Phys., 2020, 13(2), 242–252. DOI: 10.17516/1997-1397-2020-13-2-242-252.

Introduction

Systems of ordinary differential equations (ODEs) describe various dynamic processes in chemistry, physics, etc. One of the areas where ODEs may be effectively applied is the electric circuit theory. Any changes in electric circuit lead to transient processes where some voltage swells, electromagnetic oscillations, extra currents may occur. They can damage electrical devices. At the same time transient processes occur in electrical generators and other electric circuits. Many electric circuits problems are described by stiff systems of ODEs.

In some cases explicit methods are required to solve initial value problems of stiff ODEs because L-stable methods involve inversion of the Jacobi matrix of a system. This defines overall computational costs [1–2]. At the same time explicit methods do not require the Jacobi matrix computation and they are more preferable to use for problems with not so high stiffness ratio.

At present time various explicit and implicit methods were developed [3]. The former are used on transition regions where the integration step is restricted by the accuracy criterion and there is no requirements for large stability interval. The latter are for the regions where large stability interval gives an opportunity to pass the integration interval in "several steps". Nevertheless these

*mixailrybkov@yandex.ru https://orcid.org/0000-0002-6560-1435
†lvknaub@yandex.ru https://orcid.org/0000-0003-4857-2078
‡danilkhorov@gmail.com https://orcid.org/0000-0001-8967-8341

algorithms are not so efficient to solve high dimension systems of ODEs because of mentioned above reasons.

Variable order algorithms based on explicit schemes were developed [4]. They are applied to the regions where there is no need to use high-order methods. Low computational cost can be achieved by using there low-order methods with extended stability intervals which in fact play part of implicit methods from the point of view of the length of stability interval.

Low-order methods with large stability interval are needed to develop such algorithms. In addition the greater number of stages of a method (and therefore the higher the degree m of stability polynomial), the large the stability interval is. The stability polynomials of degree up to $m = 13$ were constructed [2]. The algorithm to determine the stability polynomial coefficients was developed such that the corresponding explicit Runge-Kutta methods have a predetermined shape and size of the stability region [7].

Here implementation of the algorithm to obtain the stability polynomial coefficients with the use of the library for arbitrary precision arithmetic GMP is presented. Set of the first-order methods with extended stability intervals is developed. Numerical simulation of Van der Pol oscillator shows that proposed algorithms are more efficient in comparison with the Merson method of fourth order of accuracy.

1. Explicit Runge-Kutta methods

We consider the Cauchy problem for the stiff system of ordinary differential equations

$$y' = f(t, y),\ y\big(t_0\big) = y_0,\ t_0 \leqslant t \leqslant t_k, \tag{1}$$

where y и f are real N-dimantional vector functions, t is independent variable. To solve (1) the following explicit Runge-Kutta methods were proposed [2]

$$y_{n+1} = y_n + \sum_{i=1}^{m} p_{mi} k_i, \quad k_i = hf\Big(t_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j\Big), \tag{2}$$

where k_i, $1 \leqslant i \leqslant m$, are stages of the method, h is an integration step, p_{mi}, α_i, β_{ij}, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant i-1$, are numerical coefficients that define stability and accuracy of scheme (2). For the sake of simplicity we consider the Cauchy problem for the autonomous system of ordinary differential equations

$$y' = f(y), \quad y\big(t_0\big) = y_0,\ t_0 \leqslant t \leqslant t_k. \tag{3}$$

To solve (3) we can also write formulas (2) in the following form:

$$y_{n,i} = y_n + \sum_{j=1}^{i} \beta_{i+1,j} k_j, \quad 1 \leqslant i \leqslant m-1, \quad y_{n+1} = y_n + \sum_{i=1}^{m} p_{mi} k_i, \tag{4}$$

where $k_i = hf\Big(y_{n,i-1}\Big)$, $1 \leqslant i \leqslant m$, $y_{n,0} = y_n$. The results given below can be used for non-autonomous systems if we assume in (2) that

$$\alpha_1 = 0,\ \alpha_i = \sum_{j=1}^{i-1} \beta_{ij},\ 2 \leqslant i \leqslant m. \tag{5}$$

Below we need matrix B_m with elements b_{ij} in the form [2]

$$\begin{gathered} b_{1i}=1,\ \ 1\leqslant i\leqslant m,\ \ \ b_{ki}=0,\ \ \ 2\leqslant k\leqslant m,\ \ \ 1\leqslant i\leqslant k-1, \\ b_{ki}=\sum_{j=k-1}^{i-1}\beta_{ij}b_{k-1,j},\ \ 2\leqslant k\leqslant m,\ k\leqslant i\leqslant m, \end{gathered} \tag{6}$$

where β_{ij} are numerical coefficients of schemes (2) or (4).

The stability of one-step methods is usually investigated by applying a Runge–Kutta method to a linear scalar equation known as Dahlquist's test equation

$$y'=\lambda y,\ \ \ y\big(0\big)=y_0,\ t\geqslant 0, \tag{7}$$

with complex λ, $Re\big(\lambda\big)<0$. Variable λ is considered as a certain eigenvalue of the Jacobi matrix of problems (1) or (3). Applying numerical scheme (4) to Dahlquist's equation we get

$$y_{n+1}=Q_m(z)y_n,\ z=h\lambda,\quad Q_m(z)=1+\sum_{i=1}^{m}c_{mi}z^i,\ \ \ c_{mi}=\sum_{j=1}^{m}b_{ij}p_{mj},\ \ \ 1\leqslant i\leqslant m. \tag{8}$$

Denoting $C_m=\big(c_{m1},...,c_{mm}\big)^T$ and $P_m=\big(p_{m1},...,p_{mm}\big)^T$, we can rewrite the latter equality (8) in the form

$$B_mP_m=C_m, \tag{9}$$

where the elements of matrix B_m are defined in (6). For internal numerical schemes (4) we have

$$y_{n,k}=Q_k\big(z\big)y_n,\quad Q_k\big(z\big)=1+\sum_{i=1}^{k}c_{ki}z^i,\quad c_{ki}=\sum_{j=1}^{k}b_{ij}\beta_{k+1,j},\ \ \ 1\leqslant k\leqslant m-1. \tag{10}$$

Using designations $\beta_k=\big(\beta_{k+1,1},...,\beta_{k+1,k}\big)^T$ and $c_k=\big(c_{k1},...,c_{kk}\big)^T$ we obtain that coefficients β_{ij} of internal schemes (4) and the coefficients of corresponding stability polynomials are related by the equation

$$B_k\beta_k=c_k,\ \ \ 1\leqslant k\leqslant m-1. \tag{11}$$

It follows from of (6) and (10) that $b_{ki}=c_{i-1,k-1}$, i.e., the elements of (k+1)-th column of matrix B_m are equal to coefficients of stability polynomial $Q_k\big(z\big)$. Hence, if the coefficients of the stability polynomials of the basic and intermediate numerical schemes are defined then the coefficients of methods (4) are uniquely determined from linear systems (9) and (11) with upper triangular matrices $B_i, 1\leqslant i\leqslant m$.

Expansions of the exact and approximate solutions in the Taylor series in powers of h have the form

$$\begin{gathered} y\big(t_{n+1}\big)=y\big(t_n\big)+hf+\frac{h^2}{2}f'f+O\big(h^3\big), \\ y_{n+1}=y_n+\Big(\sum_{j=1}^{m}b_{1j}p_{mj}\Big)hf+\Big(\sum_{j=2}^{m}b_{2j}p_{mj}\Big)h^2f'_nf_n+O\big(h^3\big), \end{gathered} \tag{12}$$

where the elementary differentials are taken with respect to exact $y\big(t_n\big)$ and approximate y_n solutions, respectively. Comparing relations (12) under assumption that $y\big(t_n\big)=y_n$, one can show that numerical scheme (4) has the first order of accuracy if $\sum\limits_{j=1}^{m}b_{1j}p_{mj}=1$. Hence, to design m-stage methods of the first order of accuracy it is necessary to set $c_{m1}=1$.

2. Stability polynomials

Let two integer numbers k and m, $k \leqslant m$ be given. Consider the polynomial

$$Q_{m,k}(x) = 1 + \sum_{i=1}^{k} c_i x^i + \sum_{i=k+1}^{m} c_i x^i, \tag{13}$$

where the coefficients c_i, $1 \leqslant i \leqslant k$, are given and c_i, $k+1 \leqslant i \leqslant m$, are free coefficients. The coefficients c_i, $1 \leqslant i \leqslant k$, are usually defined from the approximation requirements. Therefore, for definiteness we assume below that $c_i = 1/i!$, $1 \leqslant i \leqslant k$.

Denote extremum points of (13) as $x_1, \ldots, x_{m-1}$, where $x_1 > x_2 > \cdots > x_{m-1}$. Unknown coefficients $c_i, k+1 \leqslant i \leqslant m$ can be obtained from the condition that polynomial (13) takes on predefined values at extremum points $x_i, k \leqslant i \leqslant m-1$, i.e.,

$$Q_{m,k}\big(x_i\big) = F_i, \quad k \leqslant i \leqslant m-1, \tag{14}$$

where $F\big(x\big)$ is a preassigned function, $F_i = F\big(x_i\big)$. For this purpose let us consider the algebraic system of equations with respect to variables $x_i, k \leqslant i \leqslant m-1$, and $c_j, k+1 \leqslant j \leqslant m$,

$$Q_{m,k}\big(x_i\big) = F_i, \quad Q'_{m,k}\big(x_i\big) = 0, \quad k \leqslant i \leqslant m-1, \quad Q'_{m,k} = \sum_{i=1}^{m} i c_i x^{i-1}. \tag{15}$$

We rewrite (15) in the form that is convenient for computations. Let us introduce vectors y, z, g and r with components

$$y_i = x_{k+i-1}, \quad z_i = c_{k+i}, \quad g_i = F_{k+i-1} - 1 - \sum_{j=1}^{k} c_j y_i^j, \quad r_i = -\sum_{j=1}^{k} j c_j y_i^{j-1}, \quad 1 \leqslant i \leqslant m-k, \tag{16}$$

We also introduce diagonal matrices $E_1, \ldots, E_5$ with elements on the main diagonal

$$\begin{aligned} &e_1^{ii} = k+i, \quad e_2^{ii} = 1/y_i, \quad e_3^{ii} = \sum_{j=1}^{k} j c_j y_i^{j-1} + \sum_{j=1}^{m-k} (k+j) z_j y_i^{k+j-1}, \\ &e_4^{ii} = \sum_{j=2}^{k} j(j-1) c_j y_i^{j-2} + \sum_{j=1}^{m-k} (k+j)(k+j-1) z_j y_i^{k+j-2}, \\ &e_5^{ii} = (-1)^{k+i-1}, \quad 1 \leqslant i \leqslant m-k. \end{aligned} \tag{17}$$

Consider matrix A with elements $a^{ij} = y_i^{k+j}$, $1 \leqslant i, j \leqslant m-k$. The elements of vectors (16), matrices (17) and A depend on numbers m and k, where

$$g = g\big(y\big), \quad r = r\big(y\big), \quad E_2 = E_2\big(y\big), \quad E_3 = E_3\big(y, z\big), \quad E_4 = E_4\big(y, z\big), \quad A = A\big(y\big).$$

Then, we can rewrite problem (15) in the form

$$Az - g = 0, \quad E_2 A E_1 z - r = 0. \tag{18}$$

System (18) is ill-conditioned that leads to some difficulties in applying the fixed-point iterations for its solution. For convergence of the Newton method it is necessary to obtain good initial values but in this case it is difficult problem in its own right.

If we assume in (15) that $F_i = (-1)^i$, $k \leqslant i \leqslant m-1$, we can find the polynomial with the maximal length of stability interval. In this case the problem of finding initial value y_0 is solved by using values of the Chebyshev polynomial at extremum points over interval $[-2m^2, 0]$, where m is the degree of polynomial (13). These values are

$$y_i = m^2[\cos(i\pi/m) - 1], \quad 1 \leqslant i \leqslant m-1. \tag{19}$$

Substituting (19) in system (17), we obtain coefficients of the Chebyshev polynomial for which $\big|Q_{m1}\big(x\big)\big| \leqslant 1$ on $x \in [-2m^2, 0]$. Then for any k values given in (19) can be taken as the initial values, and according to numerical calculations there is good convergence rate in this case. If $F_i \neq (-1)^i$, $k \leqslant i \leqslant m-1$, then the choice of initial values is quite a difficult problem.

Let us describe a way to solve (18) that does not require good initial values. One can apply relaxation to solve system (18). The main idea of relaxations is to solve unsteady-state problem which solution converges to the steady-state solution of the initial problem. Let us consider the Cauchy problem

$$y' = E_5\big(E_2 A E_1 A^{-1} g - r\big), \quad y\big(0\big) = y_0. \tag{20}$$

Apparently, upon finding the stationary point of (20), the coefficients of a stability polynomial can be determined from system (18). Let us notice that because matrix E_5 is used all eigenvalues of the Jacobi matrix of (20) have negative real parts, i.e., problem (7) is stable. It follows from numerical results that (20) is a stiff problem. Methods for solving such problems use calculation of the Jacobi matrix which cause difficulties in solving (20). Therefore, we apply the method of the second order of accuracy using numerical calculation and freezing the Jacobi matrix to solve (20) [5–6].

It can be shown that values of the polynomial coefficients tend to zero as m increases. Coefficients c_i, $k+1 \leqslant i \leqslant m$, were calculated for the polynomial degree up to $m = 13$ using algorithm [2]. Moreover, the algorithm of obtaining polynomial coefficients on the interval $[-1, 1]$ was described in [7]. In this case coefficients c_i grow with slower rate and it is possible to construct polynomials of degrees $m > 13$.

3. Calculation of coefficients of stability polynomials using the GMP library

It is not difficult to see that coefficient c_m of stability polynomial (13) tends to zero as m increases and in particular if $m = 13$ and $k = 1$ the value of c_m is about 10^{-26}. Solution of problem (20) where $m > 13$ with double precision is very hard to realize because of round-off errors. In order to compute coefficients of polynomial of higher degrees m algorithm was implemented using *qd* library [8, 9].

The *qd* library allows one to perform computations with higher accuracy. Standard data type *double* which allows one to perform computations with double precision is restricted by 53 bites of binary mantissa and provides accuracy of 16 decimal digits. Whereas *qd* data type *dd_real* has 106-bit mantissa that provides accuracy of 32 decimal digits. In fact, the number of data type *dd_real* is a programmed concatenation of two *double* numbers, where mantissa becomes doubled but the range of values that can be represented using new data type stays the same (from 10^{-308} to 10^{308}). Despite this restriction accuracy of number representation increases.

With the use of this library the coefficients of polynomials up to degree $m = 35$ were computed [8]. Nevertheless, the *qd* library has some disadvantages. Firstly, accuracy of number

representation is restricted because of program implementation of data types. Secondly, it can be used only in Unix systems. Moreover, the *qd* library is written in the $C++$ programming language. That is why numerical codes that use this library could be slower than codes written in low-level programming languages (for example, C).

Here we show numerical results of calculations of polynomial coefficients with the help of the library for arbitrary precision arithmetic GMP. This library provide accuracy of computations that is restricted only by the size of random access memory. It is cross-platform library, and it supports operations on integer, rational and real numbers. Besides, the GMP library is written in the C programming language which potentially increase the speed of computations.

Using the GMP library we obtain coefficients of polynomials up to degree $m = 40$. At higher degrees there are some difficulties that may be related to the choice of initial conditions for problem (20).

4. Construction of stability regions

Let us now describe the effect of function F on the size and shape of the stability region. If we assume $F_i = (-1)^i$, $k \leqslant i \leqslant m-1$, than the stability interval length is $\left|\gamma_m\right| = 2m^2$. In this case, we have the maximum length of stability interval along the real axis for given m. The stability region of such methods is almost multiply connected which leads to the reduction of stability interval length because rounding errors may cause small imaginary parts of Jacobi matrix eigenvalues to appear. Fig. 1 shows the stability region of 5-stage method, where the stability interval length is $\left|\gamma_m\right| = 50$.

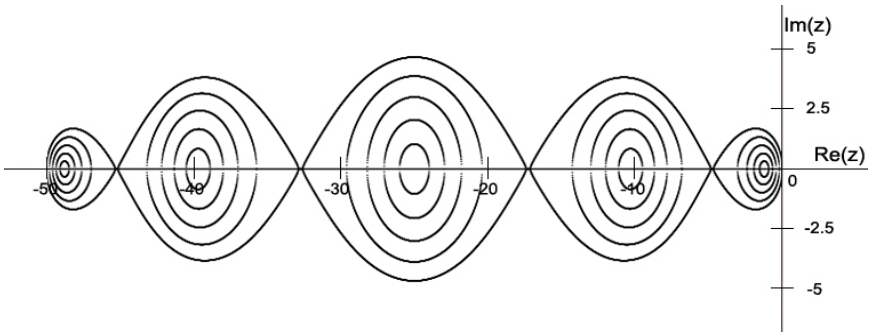

Fig. 1. Stability region at $m = 5$, $k = 1$, $F = \{-1, 1, -1, 1\}$, $\left|\gamma_m\right| = 50$

In order to avoid the stability region reduction because of rounding errors it should be "stretched" along the imaginary axis at the extremum points of the stability polynomial. For that we can assume $F_i = (-1)^i\mu$, $1 \leqslant i \leqslant m-1$, $0 < \mu < 1$. For example, if we choose $\mu = 0.95$ then the stability interval length is reduced by only 3–4% in comparison with the maximal possible length that is equal to $2m^2$. Then it becomes equal to $\left|\gamma_m\right| = 48.39$ (Fig. 2). The stability region of the 5-stage method at $\mu = 0.8$ is shown in Fig. 3. In this situation, the stability interval length is reduced to $\left|\gamma_m\right| = 43.55$ with conjoined "stretching" along the imaginary axis. For better visualization of the roots of polynomial (13) level lines $\left|Q_{m,k}(x)\right| = 1$, $\left|Q_{m,k}(x)\right| = 0.8$, $\left|Q_{m,k}(x)\right| = 0.6$, $\left|Q_{m,k}(x)\right| = 0.4$, $\left|Q_{m,k}(x)\right| = 0.2$. in the complex plane Are shown in all figures.

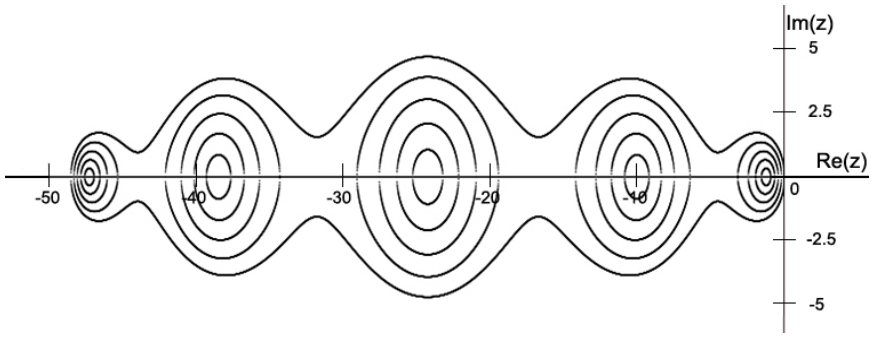

Fig. 2. Stability region at $m = 5$, $k = 1$, $F = \{-0.95, 0.95, 0.95, 0.95\}$, $\left|\gamma_m\right| = 48.39$

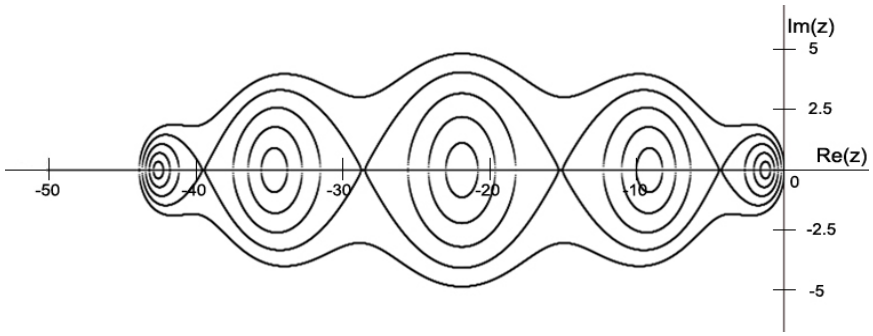

Fig. 3. Stability region at $m = 5$, $k = 1$, $F = \{-0.8, 0.8, -0.8, 0.8\}$, $\left|\gamma_m\right| = 43.55$

5. First-order method

For numerical solution of Cauchy problem (1) we consider the explicit five-stage Runge-Kutta method

$$\begin{aligned} y_{n+1} &= y_n + p_1k_1 + p_2k_2 + p_3k_3 + p_4k_4 + p_5k_5, \\ k_1 &= hf\big(y_n\big), \quad k_2 = hf\big(y_n + \beta_{21}k_1\big), \\ k_3 &= hf\big(y_n + \beta_{31}k_1 + \beta_{32}k_2\big), \\ k_4 &= hf\big(y_n + \beta_{41}k_1 + \beta_{42}k_2 + \beta_{43}k_3\big), \\ k_5 &= hf\big(y_n + \beta_{51}k_1 + \beta_{52}k_2 + \beta_{53}k_3 + \beta_{54}k_4\big), \end{aligned} \tag{21}$$

where y and f are real N-dimensional vector functions, t is independent variable, h is the integration step, k_1, k_2, k_3, k_4 and k_5 are stages of the method, $p_1, p_2, p_3, p_4, p_5, \beta_{21}, \beta_{31}, \beta_{32}$, $\beta_{41}, \beta_{42}, \beta_{43}, \beta_{51}, \beta_{52}, \beta_{53}, \beta_{54}$ are numerical coefficients that define accuracy and stability of (21). Applying the algorithm, we obtain coefficients of the stability polynomial:

$$c_{5,1} = 0.1e1, \quad c_{5,2} = 0.164341322127140896342e0, \quad c_{5,3} = 0.948975952580473808808e - 2,$$
$$c_{5,4} = 0.223956930863224544258e - 3, \quad c_{5,5} = 0.185097275222353341 53e - 5$$

In this case the stability interval length is $\left|\gamma_m\right| = 48.39$. Upon solving (9) and (11), we obtain the coefficients of the first-order method:

$\beta_{21} = 0.0413243016210550, \quad \beta_{31} = 0.0805823881610573, \quad \beta_{32} = 0.0805823881610573,$

$\beta_{41} = 0.11916681511228434, \quad \beta_{42} = 0.1597820013984078, \quad \beta_{43} = 0.0819394878966193,$

$\beta_{51} = 0.1570787892802991, \quad \beta_{52} = 0.2379583021959820, \quad \beta_{53} = 0.1631711307360486,$

$\beta_{54} = 0.0822916178203657,$

$p_1 = 0.1945277188657676, \quad p_2 = 0.3151822878089125, \quad p_3 = 0.2437005934695969,$

$p_4 = 0.1641555613805598, \quad p_5 = 0.0824338384751631.$

Accuracy control of numerical scheme is based on local error estimation [10].

The magnitude of

$$A'_n = [(0.5 - c_{m2})/\alpha_2](k_2 - k_1) \tag{22}$$

is used as preliminary estimation of local error. To we estimate the final accuracy the magnitude of

$$A''_n = (0.5 - c_{m2})(hf\big(y_{n+1}\big) - k_1) \tag{23}$$

is used. Thus, the following inequalities

$$A'_n \leqslant \epsilon, \quad A''_n \leqslant \epsilon. \tag{24}$$

are used for the accuracy control and for the choice of of integration step. As k_1 linearly depends on integration step then omission of inequality (24) leads to just one additional computation of the right hand side of the problem. If the step of integration is successful the second inequality (24) does not lead to the increase of computational cost because $f\big(y_{n+1}\big)$ is not used at the next step. At the same time if the second inequality (24) is used for accuracy control the repeat computations in the case of violating accuracy criterion are quite expensive. Moreover, the greater m the higher computational cost is. Nevertheless, in most cases preliminary estimation of A'_n allows one to avoid repeat computations. The following inequality

$$\nu_n \leqslant \gamma_{m,1} \tag{25}$$

is used for stability control of method (2), where

$$\nu_n = \left|\alpha_2\beta_{32}\right|^{-1} \max_{1 \leqslant j \leqslant N} \left|[\alpha_2 k_3 + \alpha_3 k_2 - (\alpha_2 + \alpha_3)k_1]_j / [k_2 - k_1]_j\right|, \tag{26}$$

and positive constants $\gamma_{m,1}$ define the size of stability regions [10].

6. Merson method

One of the most effective explicit fourth-order Runge-Kutta type methods is the Merson method [8]

$$\begin{aligned} &y_{n+1} = y_n + \frac{1}{6}k_1 + \frac{2}{3}k_4 + \frac{1}{6}k_5, \\ &k_1 = hf\big(y_n\big), \quad k_2 = hf\big(y_n + \frac{1}{3}k_1\big), \quad k_3 = hf\big(y_n + \frac{1}{6}k_1 + \frac{1}{6}k_2\big), \\ &k_4 = hf\big(y_n + \frac{1}{8}k_1 + \frac{3}{8}k_3\big), \quad k_5 = hf\big(y_n + \frac{1}{2}k_1 - \frac{3}{2}k_3 + 2k_4\big). \end{aligned} \tag{27}$$

The fifth computation of function f does not result in the fifth order of accuracy but allows one to extend the stability interval to 3.5 and estimate truncation error $\delta_{n,4}$ using stages k_i i.e.,

$$\delta_{n,4} = (2k_1 - 9k_3 + 8k_4 - 2k_5)/30.$$

We use inequality $\|\delta_{n,4}\| \leqslant 5\epsilon^{5/4}$ for accuracy control. Despite the fact that inequality for accuracy control is obtained with the help of a linear equation, it shows high reliability in solving non-linear problems.

Now let us construct the inequality for stability control. Applying to $k_3 - k_2$ the first order Taylor's formula with the remainder term written in the Lagrangian form, we have

$$k_3 - k_2 = h[\partial f(\mu_n)/\partial y](k_2 - k_1)/6,$$

where vector μ_n is calculated in some vicinity of solution $y(t_n)$. Taking into account that

$$k_2 - k_1 = h^2 f'_n f_n/3 + O(h^3),$$

the inequality

$$\nu_{n,4} = 6 \cdot \max_{1 \leqslant j \leqslant N} \left| \frac{k_3^j - k_2^j}{k_2^j - k_1^j} \right| \leqslant 3.5 \tag{28}$$

can be used for stability control of (27), where 3.5 is the approximate length of stability interval. Let $\epsilon_{n,4} = \delta_{n,4}/5$. Then inequalities $\epsilon_{n,4} \leqslant 5\epsilon^{5/4}$ and $\nu_{n,4} \leqslant 3.5$. can be used for accuracy and stability control of scheme (27), respectively.

7. Numerical results

The computations were performed on Intel(R) Core(TM) i7-8550U CPU. However, coefficients of stability polynomial were computed with the help of the GMP library whereas solution of differential problem was determined with double precision. The norm $\|\xi_n\|$ in the inequalities for the accuracy control was calculated by the formula

$$\|\xi_n\| = \max_{1 \leqslant j \leqslant N} |\xi_n^i|/(|y_n^i| + r),$$

where i is a number of vector component, r is a positive parameter. If inequality $|y_n^i| < r$ is satisfied for the component with number i then the absolute error $r \cdot \varepsilon$ is controlled. Otherwise we control the relative error ε, where ε is the required accuracy.

We chose the Van der Pol oscillator (29) as a test example. This problem has the stiffness ratio approximately equal to 10^6:

$$\begin{aligned} &y_1' = y_2, \quad y_2' = \Big((1 - y_1^2)y_2 - y_1\Big)/10^{-6}, \\ &0 \leqslant t \leqslant 1, \quad h_0 = 10^{-3}, \quad y_1(0) = 2, \quad y_2(0) = 0, \quad \varepsilon = 10^{-2}. \end{aligned} \tag{29}$$

The efficiency of two algorithms are compared. The first algorithm is the first order 5-stage Runge-Kutta method described in Section 5. The second algorithm is the traditional 5-stage Merson method of the forth order of accuracy (27). Both algorithms were applied in two modes: with stability control and without it. We counted total numbers of steps, repeat computations of a solution (due to omission of the defined accuracy), and the number computations of the

right hand side the problem. The accuracy $\varepsilon = 10^{-2}$ was supported with the Merson method, whereas for the first-order method we needed to use $\varepsilon = 10^{-5}$ in order to provide 10^{-2} in fact. Nevertheless, under these conditions the constructed algorithm shows better efficiency (Fig. 4).

The comparison of two algorithms shows that stability control increases efficiency because extra repeat computations of solution originating from instability of numerical scheme are eliminated. In addition, the constructed first-order algorithm has less computational cost estimated by the number of computations of the right hand side. Simulations of other test examples show similar pattern.

Criterion	First-order method without stability control	First-order method with stability control	Merson method without stability control	Merson method with stability control
Number of integration steps	69 433	51 414	549 241	556 114
Number of repeated solution calculations	20 001	1 052	187 120	6 464
Number of right part computations	452 683	309 948	3 494 685	2 806 426

Fig. 4. Numerical results for the Van der Pol oscillator problem

Conclusion

Implementation of the algorithm to obtain coefficients of stability polynomial with the use of the GMP library allowed one to build stability polynomial up to degree $m = 40$. It provides a possibility to develop methods with extended stability regions with respective number of stages. The greater number of stages the larger stability interval is, and therefore the higher efficiency of numerical scheme is achieved in the case of stiff problems.

Comparing two five-stage methods (the proposed first-order method and the Merson method), one can see that at the same number of stages extending the stability interval decreases the overall computational cost.

It is important to say that the first-order methods with extended stability regions allow one to significantly increase the efficiency in the region where the step is restricted by stability. So methods described here can be used in adaptive algorithms where number of stages may vary from one integration step to another. It provides large stability interval where it is needed and decreases computational cost when numerical scheme is unconditionally stable.

The study was funded by Russian Foundation for Basic Research (project no. 18-31-00375).

References

[1] E.Hairer, G.Wanner, Solving ordinary differential equations II. Stiff and differential-algebraic problems, Berlin, Springer, 1996.

[2] E.A.Novikov, Explicit methods for stiff systems, Novosibirsk, Nauka, 1997 (in Russian).

[3] E.A.Novikov, A.E.Novikov, Explicit-Implicit Variable Structure Algorithm for Solving Stiff Systems, *International Journal of Mathematical Models and Methods in Applied Sciences*, **9**(2015), no. 1, 62–70.

[4] E.A.Novikov, Yu.V.Shornikov, Computer simulation of stiff hybrid systems, Novosibirsk: Publisher of NSTU, 2012 (in Russian).

[5] A.E.Novikov, E.A.Novikov, L-stable (2,1)-method for stiff nonautonomius problem solving, *Computing technologies*, **13**(2008), 477–482 (in Russian).

[6] E.A.Novikov, Yu.A.Shitov, Integration algorithm for stiff systems based on a second-order accuracy (m, k)–method with numerical calculation of the Jacobi matrix, Krasnoyarsk: Preprint of the Exhibition Center of the Siberian Branch of the USSR Academy of Sciences no. 20, 1988 (in Russian).

[7] E.A.Novikov, M.V.Rybkov, The numerical algorithm of constructing stability polynomials of first order methods, *Bulletin of the Buryat State University*, (2014), no. 9–2, 80–85 (in Russian).

[8] E.A.Novikov, M.V.Rybkov, The numerical algorithm of constructing of stability regions for explicit methods, *Control systems and information technologies*, **55**(2014), no. 1.1, 173–177 (in Russian).

[9] Yozo Hida, Xiaoye S Li, David H Bailey, Quad-double arithmetic: algorithms, implementation, and application, Technical Report LBNL–46996, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, 2000.

[10] L.V.Knaub, P.S.Litvinov, A.E.Novikov, M.V.Rybkov, Solving Problems of Moderate Stiffness Using Methods of the First Order with Conformed Stability Domains, *University Scientific Journal*, (2016), no. 22, 49–58.

[11] R.H.Merson, An operational methods for integration processes, Proc. of Symp. on Data Processing, Weapons Research Establishment, Salisbury, Australia, 1957.

Методы первого порядка с расширенными областями устойчивости для расчета задач электрических цепей

Михаил В. Рыбков
Людмила В. Кнауб
Данил В. Хоров
Сибирский федеральный университет
Красноярск, Российская Федерация

Аннотация. Исследуется применение контроля устойчивости численных схем типа Рунге-Кутты для повышения эффективности при интегрировании жестких задач. Приведена реализация алгоритма определения коэффициентов полиномов устойчивости, при которых метод имеет заданную форму и размер области устойчивости, с помощью библиотеки GMP. Построены наборы методов первого порядка с расширенными областями устойчивости. Приведены результаты расчетов задач из теории электрических цепей, показывающие повышение эффективности построенных методов первого порядка точности в сравнении с методом более высокого порядка.

Ключевые слова: жесткая задача, явные методы, контроль точности, контроль устойчивости.