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

Журнал Сибирского федерального университета

Математика и физика

Journal of Siberian Federal University

Mathematics & Physics

2020 13 (6)

ISSN 1997-1997-1397
(Print)

ISSN 2313-6022
(Online)

2020 13 (6)

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
№ 6. 26.12.2020. 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 15.12.20. Формат 84×108/16. Усл.печ. л. 11,9.
Уч.-изд. л. 11,6. Бумага тип. Печать офсетная.
Тираж 1000 экз. Заказ 11993
Возрастная маркировка в соответствии с Федеральным законом № 436-ФЗ:16+

CONTENTS

СОДЕРЖАНИЕ

DOI: 10.17516/1997-1397-2020-13-6-661–669
УДК 517.977.55:536.25

On a Creeping 3D Convective Motion of Fluids with an Isothermal Interface

Viktor K. Andreev*
Institute of Computational Modelling SB RAS
Krasnoyarsk, Russian Federation
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 22.06.2020, received in revised form 02.07.2020, accepted 20.09.2020

Abstract. In the work the 3D two-layer motion of liquids, the velocity field of which has a special form, is considered. The arising conjugate initial boundary value problem for the Oberbek–Boussinesq model is reduced to a system of ten integrodifferential equations with full conditions on a flat interface. It is shown that for small Marangoni numbers the stationary problem can have up to two solutions. The case when the stationary flow arises due to a change in the internal interphase energy is analyzed separately.

Keywords: Oberbek-Boussinesq model, interphase energy, creeping flow, inverse problem.

Citation: V.K. Andreev, On a Creeping 3D Convective Motion of Fluids with an Isothermal Interface, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 661–669. DOI: 10.17516/1997-1397-2020-13-6-661-669.

1. Statement of the problem and basic equations

Suppose that two viscous heat-conducting fluids with a common interface $z = l_1 < l_2$ move in a layer $|x| < \infty$, $|y| < \infty$, $0 < z < l_2$, l_j are constants. The fluid " 1 " occupies the region $0 < z < l_1$ and fluid " 2 " occupies the region $l_1 < z < l_2$. The planes $z = 0$ and $z = l_2$ are solid fixed walls, the force of gravity is directed perpendicular to the layers. Oberbeck-Boussinesq equations are used as a mathematical model of fluid motion. Solutions are sought in a special way

$$u_j(x,y) = (f_j(z,t) + h_j(z,t))x, \quad v_j(x,y) = (f_j(z,t) - h_j(z,t))y, \quad w_j = -2\int_{z_0}^{z} f_j(\xi,t)\,d\xi, \quad (1)$$

$$\frac{1}{\rho_j}p_j = b_j(z,t)x^2 + d_j(z,t)y^2 + q_j(z,t), \quad (2)$$

$$T_j = a_j(z,t)x^2 + c_j(z,t)y^2 + \theta_j(z,t), \quad (3)$$

where $u_j(x,y,z,t)$, $v_j(x,y,z,t)$, $w_j(x,y,z,t)$ are projections of velocity vectors on the x, y, z axis, respectively; $p_j(x,y,z,t)$ are pressures; ρ_j are constants of density; $T_j(x,y,z,t)$ are absolute temperatures, $j = 1, 2$. The functions f_j, h_j, b_j, d_j, q_j, a_j, c_j, θ_j are new unknown function.

Substitution of the formulas (1)–(3) in the systems of Oberbeck-Boussinesq equations leads to the following systems

$$f_{jt} + f_j^2 + h_j^2 - 2f_{jz}\int_{z_0}^{z} f_j(\xi,t)\,d\xi + g\beta_j\int_{z_0}^{z}(a_j(\xi,t) + c_j(\xi,t))\,d\xi = \nu_j f_{jzz} + n_{j1}(t), \quad (4)$$

*andr@icm.krasn.ru

$$h_{jt} + 2f_j h_j - 2h_{jz}\int_{z_0}^{z} f_j(\xi,t)\,d\xi + g\beta_j\int_{z_0}^{z}\left(a_j(\xi,t) - c_j(\xi,t)\right)\,d\xi = \nu_j h_{jzz} + n_{j2}(t), \tag{5}$$

$$a_{jt} + 2a_j(f_j + h_j) - 2a_{jz}\int_{z_0}^{z} f_j(\xi,t)\,d\xi = \chi_j a_{jzz}, \tag{6}$$

$$c_{jt} + 2c_j(f_j - h_j) - 2c_{jz}\int_{z_0}^{z} f_j(\xi,t)\,d\xi = \chi_j c_{jzz}, \tag{7}$$

$$\theta_{jt} - 2\theta_{jz}\int_{z_0}^{z} f_j(\xi,t)\,d\xi = \chi_j\theta_{jzz} + 2\chi_j(a_j + c_j). \tag{8}$$

Here $\nu_j > 0$, $\chi_j > 0$, $\beta_j > 0$ are constants of kinematic viscosities, thermal diffusivities and thermal expansion coefficients of liquids; $n_{j1}(t)$, $n_{j2}(t)$ are arbitrary functions of time. By the known functions a_j, c_j the functions b_j, d_j are determined by quadratures

$$b_j(z,t) = g\beta_j\int_{z_0}^{z} a_j(\xi,t)\,d\xi - n_{j1}(t), \quad d_j(z,t) = g\beta_j\int_{z_0}^{z} c_j(\xi,t)\,d\xi - n_{j2}(t). \tag{9}$$

In the integral terms, the constant z_0 is equal to "0" for the first fluid ($j = 1$) and l_1 for the second fluid ($j = 2$). It can be verified that pressures in liquids are determined as follows

$$\begin{aligned}\frac{1}{\rho_j}p_j = &\left[g\beta_j\int_{z_0}^{z} a_j(\xi,t)\,d\xi - n_{j1}(t)\right]x^2 + \left[g\beta_j\int_{z_0}^{z} c_j(\xi,t)\,d\xi - n_{j2}(t)\right]y^2 - 2\nu_j f_j - gz + \\ &+ g\beta_j\int_{z_0}^{z}\theta_j(\xi,t)\,d\xi + 2\int_{z_0}^{z}(z-\xi)f_{jt}(\xi,t)\,d\xi + 2\left(\int_{z_0}^{z} f(\xi,t)\,d\xi\right)^2 + q_{j0}(t),\end{aligned} \tag{10}$$

with arbitrary functions $q_{j0}(t)$.

Remark 1. The velocity field (1), proposed in [1] is a special case of the velocity field for the Navier-Stokes equations [2].

2. Boundary and initial conditions

On solid boundaries, the sticking conditions for the velocities are satisfied, which implies equalities

$$f_1(0,t) = h_1(0,t) = 0, \quad f_2(l_2,t) = h_2(l_2,t) = \int_{l_1}^{l_2} f_2(\xi,t)\,d\xi = 0 \tag{11}$$

And the temperature is set

$$\begin{aligned}a_1(0,t) = a_0(t), \quad c_1(0,t) = c_0(t), \quad \theta_1(0,t) = \theta_1(t),\\ a_2(l_2,t) = a_2(t), \quad c_2(l_2,t) = c_2(t), \quad \theta_2(l_2,t) = \theta_2(t).\end{aligned} \tag{12}$$

The top wall can also be thermally insulated

$$a_{2z}(l_2,t) = c_{2z}(l_2,t) = \theta_{2z}(l_2,t) = 0, \tag{13}$$

To formulate the conditions on the undeformed interface $z = l_1$, we assume that the surface tension depends linearly on temperature

$$\sigma(T) = \sigma_0 - æ(T - T_0), \tag{14}$$

whereσ_0, æ, T_0 are given positive constants, $T(x, y, l_1, t)$ is temperature on this border.

On the interface $z = l_1$ there are equalities of velocities and temperatures. Taking into account the representation (1), (3) we get [3]

$$\begin{gathered} f_1(l_1,t) = f_2(l_1,t), \quad h_1(l_1,t) = h_2(l_1,t), \quad a_1(l_1,t) = a_2(l_1,t), \\ c_1(l_1,t) = c_2(l_1,t), \quad \theta_1(l_1,t) = \theta_2(l_1,t). \end{gathered} \tag{15}$$

Tangential stresses are reduced to two relations

$$\begin{aligned} \mu_2 f_{2z}(l_1,t) - \mu_1 f_{1z}(l_1,t) &= -æ(a_1(l_1,t) + c_1(l_1,t)), \\ \mu_2 h_{2z}(l_1,t) - \mu_1 h_{1z}(l_1,t) &= -æ(a_1(l_1,t) - c_1(l_1,t)), \end{aligned} \tag{16}$$

where $\mu_j = \rho_j \nu_j$ are dynamic viscosity of liquids.

The kinematic condition for a fixed and non-deformable interface ($w_1(l_1,t) = w_2(l_1,t) = 0$) is equivalent to the integral equality

$$\int_0^{l_1} f_1(\xi,t)\, d\xi = 0. \tag{17}$$

The energy condition [3], taking into account the assumptions (8), can be written as

$$k_2 T_{2z}(x,y,l_1,t) - k_1 T_{1z}(x,y,l_1,t) = æ T(x,y,l_1,t) \mathrm{div}_\Gamma \mathbf{u}. \tag{18}$$

where k_j are constant coefficients of thermal conductivity of liquids; $\mathrm{div}_\Gamma \mathbf{u}$ is surface divergence of the velocity vector; $T(x,y,l_1,t) = T_1(x,y,l_1,t) = T_2(x,y,l_1,t)$. Since in our case $\mathrm{div}_\Gamma \mathbf{u} = u_x + v_y$, then using the formulas (1), (3) from (18) we derive the relations

$$\begin{aligned} k_2 a_{2z}(l_1,t) - k_1 a_{1z}(l_1,t) &= 2æ a_1(l_1,t) f_1(l_1,t), \\ k_2 c_{2z}(l_1,t) - k_1 c_{1z}(l_1,t) &= 2æ c_1(l_1,t) f_1(l_1,t), \\ k_2 \theta_{2z}(l_1,t) - k_1 \theta_{1z}(l_1,t) &= 2æ \theta_1(l_1,t) f_1(l_1,t). \end{aligned} \tag{19}$$

The relation order of equation right-hand side (18) to the first terms of its left-hand side is estimated by the parameter $E = æ^2 \theta^* / \mu_2 k_2$ (for the second term $\mu_1 k_1$), where θ^* is the characteristic temperature on the interface [3]. These parameters for ordinary liquid media are small and instead of (18) the equality of heat fluxes is used. However, for low-viscosity liquids and small k_j the right-hand side in (18) (right-hand sides in (19)) must be taken into account, for example, for cryogenic media [3].

At the initial moment of time, all functions are set

$$f_j(z,0) = f_{j0(z)}, \quad h_j(z,0) = h_{j0}(z), \quad a_j(z,0) = a_{j0}(z), \quad c_j(z,0) = c_{j0}(z), \quad \theta_j(z,0) = \theta_j 0(z), \tag{20}$$

that satisfy the conditions of agreement with (12), (13), (15)–(17), (19). For example, $f_{10}(l_1) = f_{20}(l_1)$ etc.

Remark 2. The formulated initial-boundary value problem (4)–(9), (11)–(17), (19), (20) is the inverse, since the functions $n_{j1}(t)$, $n_{j2}(t)$ must be found along with its solution. For a complete statement of this problem, two more conditions must be set

$$\int_0^{l_1} h_1(\xi,t)\, d\xi = 0, \quad \int_{l_1}^{l_2} h_2(\xi,t)\, d\xi = 0, \tag{21}$$

which together with the integral equalities (11), (17) mean closedness of motion.

3. Dimensionless variables

We introduce dimensionless variables and parameters

$$\begin{gathered}\tau=\frac{\chi_1}{l_1^2}t,\quad \xi=\frac{z}{l_2},\quad \chi=\frac{\chi_1}{\chi_2},\quad \mathrm{P}_j=\frac{\nu_j}{\chi_j},\quad \mu=\frac{\mu_1}{\mu_2},\quad k=\frac{k_1}{k_2},\quad l=\frac{l_1}{l_2}<1,\\ \mathrm{G}_j=\frac{a^*l_2l_1^4g\beta_j}{\chi_1^2},\quad \varepsilon_j=\frac{\chi_j}{\chi_1},\quad \mathrm{M}=\frac{æa^*l_1^2l_2}{\mu_2\chi_1},\quad F_j(\xi,\tau)=\frac{l_1^2}{\chi_1\mathrm{M}}f_j(z,t),\\ H_j(\xi,\tau)=\frac{l_1^2}{\chi_1\mathrm{M}}h_j(z,t),\quad A_j(\xi,\tau)=\frac{a_j(z,t)}{a^*\mathrm{M}},\quad C_j(\xi,\tau)=\frac{c_j(z,t)}{a^*\mathrm{M}},\\ N_j(\tau)=\frac{l_1^4n_j(t)}{\chi_1^2\mathrm{M}},\quad Q_j(\xi,\tau)=\frac{\theta_j(z,t)}{a^*l_1^2\mathrm{M}}.\end{gathered}\tag{22}$$

Here P_j are Prandtl numbers, G_j are Grashof numbers, M is Marangoni number. It is further believed that $a^*=\max\limits_{t\geqslant0}|a_1(t)|>0$ and the characteristic temperature at the interface is $\theta^*=a^*l_1^2$.

In the new variables, the system (4)–(8) will be rewritten as follows

$$F_{j\tau}+\mathrm{M}\Big[F_j^2+H_j^2-2F_{j\xi}\int_{z_0/l_2}^{\xi}F_j(\zeta,\tau)\,d\zeta\Big]+\mathrm{G}_j\int_{z_0/l_2}^{\xi}(A_j(\zeta,\tau)+C_j(\zeta,\tau))\,d\zeta=\\ =\mathrm{P}_jl^2\varepsilon_jF_{j\xi\xi}+N_{j1}(\tau),\tag{23}$$

$$H_{j\tau}-2\mathrm{M}\Big[F_jH_j-2H_{j\xi}\int_{z_0/l_2}^{\xi}F_j(\zeta,\tau)\,d\zeta\Big]+\mathrm{G}_j\int_{z_0/l_2}^{\xi}(A_j(\zeta,\tau)-C_j(\zeta,\tau))\,d\zeta=\\ =\mathrm{P}_jl^2\varepsilon_jH_{j\xi\xi}+N_{j2}(\tau),\tag{24}$$

$$A_{j\tau}+2\mathrm{M}A_j(F_j+H_j)-2\mathrm{M}A_{j\xi}\int_{z_0/l_2}^{\xi}F_j(\zeta,\tau)\,d\zeta=l^2\varepsilon_jA_{j\xi\xi},\tag{25}$$

$$C_{j\tau}+2\mathrm{M}C_j(F_j-H_j)-2\mathrm{M}C_{j\xi}\int_{z_0/l_2}^{\xi}F_j(\zeta,\tau)\,d\zeta=l^2\varepsilon_jC_{j\xi\xi},\tag{26}$$

$$Q_{j\tau}-2\mathrm{M}Q_{j\xi}\int_{z_0/l_2}^{\xi}F_j(\zeta,\tau)\,d\zeta=l^2\varepsilon_jQ_{j\xi\xi}+2\varepsilon_j(A_j+C_j).\tag{27}$$

In integral expressions for $j=1$ the $z_0=0$ and at $j=2$ we have $z_0=l_1$, so that $0<\xi<l$ in the first layer and $l<\xi<1$ in the second layer.

The boundary conditions (11)–(13), (15)–(17), (19), (21) are rewritten as

$$F_1(0,\tau)=H_1(0,\tau)=0,\quad F_2(1,\tau)=H_2(1,\tau)=\int_l^1F_2(\xi,\tau)\,d\xi=0,\tag{28}$$

$$\begin{gathered}A_1(0,\tau)=A_1(\tau),\quad C_1(0,\tau)=C_1(\tau),\quad Q_1(0,\tau)=Q_1(\tau),\\ A_2(1,\tau)=A_2(\tau),\quad C_2(1,\tau)=C_2(\tau),\quad Q_2(1,\tau)=Q_2(\tau),\end{gathered}\tag{29}$$

$$A_{2\xi}(1,\tau)=C_{2\xi}(1,\tau)=Q_{2\xi}(1,\tau)=0,\tag{30}$$

$$\begin{gathered}F_1(l,\tau)=F_2(l,\tau),\quad H_1(l,\tau)=H_2(l,\tau),\quad A_1(l,\tau)=A_2(l,\tau),\\ C_1(l,\tau)=C_2(l,\tau),\quad Q_1(l,\tau)=Q_2(l,\tau),\end{gathered}\tag{31}$$

$$\begin{gathered}F_{2\xi}(l,\tau)-\mu F_{1\xi}(l,\tau)=-\mathrm{M}(A_1(l,\tau)+C_1(l,\tau)),\\ H_{2\xi}(l,\tau)-\mu H_{1\xi}(l,\tau)=-\mathrm{M}(A_1(l,\tau)-C_1(l,\tau)),\end{gathered}\tag{32}$$

$$\int_0^l F_1(\xi,\tau)\,d\xi = 0, \tag{33}$$

$$\begin{aligned} A_{2\xi}(l,\tau) - kA_{1\xi}(l,\tau) &= 2\mathrm{E}A_1(l,\tau)F_1(l,\tau),\\ C_{2\xi}(l,\tau) - kC_{1\xi}(l,\tau) &= 2\mathrm{E}C_1(l,\tau)F_1(l,\tau),\\ Q_{2\xi}(l,\tau) - kQ_{1\xi}(l,\tau) &= 2\mathrm{E}Q_1(l,\tau)F_1(l,\tau), \end{aligned} \tag{34}$$

$$\int_0^l H_1(\xi,\tau)\,d\xi = 0, \quad \int_l^1 H_2(\xi,\tau)\,d\xi = 0. \tag{35}$$

The initial data (20) will be of the form

$$\begin{gathered} F_j(\xi,0) = F_{j0}(\xi), \quad H_j(\xi,0) = H_{j0}(\xi), \quad A_j(\xi,0) = A_{j0}(\xi),\\ C_j(\xi,0) = C_{j0}(\xi), \quad Q_j(\xi,0) = Q_{j0}(\xi). \end{gathered} \tag{36}$$

4. Stationary creeping flow with an isothermal interface

In this case, the right-hand sides (32) must be zero. It means that $A_1(l,\tau) = C_1(l,\tau) = 0$ and the task set above will be redefined. Here we consider the creeping motion ($\mathrm{M} \ll 1$). It is necessary to assume that the initial initial data are of the order M. Let $\mathrm{M} \to 0$, then the equations (23)–(27) will be linear and the right-hand sides of the boundary conditions are equal to zero. However, the relations (34) remain nonlinear.

Remark 3. If, assume that $A_j(\xi,\tau) = 0$, $C_j(\xi,\tau) = 0$, then the interface will be isothermal: $T_1(x,y,l,\tau) = T_2(x,y,l,\tau) = \theta_1(l,\tau) = \theta_2(l,\tau) = 0$.

In this paragraph, we assume that the upper plane is thermally insulated and conditions (30) are satisfied on it; initial data (36) are omitted. Let A_1^s, C_1^s, Q_1^s are specified stationary values of boundary conditions (29). Not complicated, but rather long calculations lead to representations

$$\begin{gathered} A_1(\xi) = \alpha_1\xi + A_1^s, \quad A_2(\xi) = \alpha_2 \equiv \alpha_1 l + A_1^s,\\ C_1(\xi) = \gamma_1\xi + C_1^s, \quad C_2(\xi) = \gamma_2 \equiv \gamma_1 l + C_1^s,\\ F_1(\xi) = \frac{1}{\mathrm{P}_1 l^2}\left[\mathrm{G}_1\left(\frac{\alpha_1+\gamma_1}{24}\xi^4 + \frac{A_1^s + C_1^s}{6}\xi^3\right) - \frac{N_{11}\xi^2}{2}\right] + D_1\xi,\\ F_2(\xi) = \frac{\chi}{\mathrm{P}_2 l^2}\left[\mathrm{G}_2(\alpha_2+\gamma_2)\left(\frac{\xi^3-1}{6} - \frac{l}{2}(\xi^2-1)\right) - \frac{N_{21}}{2}(\xi^2-1)\right] + D_2(\xi-1), \end{gathered} \tag{37}$$

$$\begin{aligned} H_1(\xi) &= \frac{1}{\mathrm{P}_1 l^2}\left[\mathrm{G}_1\left(\frac{\alpha_1-\gamma_1}{24}\xi^4 + \frac{A_1^s - C_1^s}{6}\xi^3\right) - \frac{N_{12}\xi^2}{2}\right] + D_3\xi,\\ H_2(\xi) &= \frac{\chi}{\mathrm{P}_2 l^2}\left[\mathrm{G}_2(\alpha_2-\gamma_2)\left(\frac{\xi^3-1}{6} - \frac{l}{2}(\xi^2-1)\right) - \frac{N_{22}}{2}(\xi^2-1)\right] + D_4(\xi-1). \end{aligned} \tag{38}$$

The constants $D_1,\ldots,D_4$ are found from the integral equalities (28), (33), (35):

$$\begin{aligned} D_1 &= \frac{1}{3\mathrm{P}_1 l}\left[N_{11} - \mathrm{G}_1\left(\frac{(\alpha_1+\gamma_1)l^2}{20} + \frac{(A_1^s + C_1^s)l}{4}\right)\right],\\ D_2 &= \frac{\chi}{\mathrm{P}_2 l^2}\left[\frac{N_{21}(l+2)}{3} + \frac{\mathrm{G}_2(\alpha_2+\gamma_2)(l^2+2l-1)}{4}\right],\\ D_3 &= \frac{1}{3\mathrm{P}_1 l}\left[N_{12} - \mathrm{G}_1\left(\frac{(\alpha_1-\gamma_1)l^2}{20} + \frac{(A_1^s - C_1^s)l}{4}\right)\right],\\ D_4 &= \frac{\chi}{\mathrm{P}_2 l^2}\left[\frac{N_{22}(l+2)}{3} + \frac{\mathrm{G}_2(\alpha_2-\gamma_2)(l^2+2l-1)}{4}\right]. \end{aligned} \tag{39}$$

By virtue of (37),

$$\alpha_2+\gamma_2=(\alpha_1+\gamma_1)l+A_1^s+C_1^s,\quad \alpha_2-\gamma_2=(\alpha_1-\gamma_1)l+A_1^s-C_2^s. \tag{40}$$

To determine the remaining unknowns $\alpha_1,\gamma_1,N_{11},N_{21},N_{12},N_{22}$, there are relations

$$\begin{gathered}F_1(l)=F_2(l),\quad F_{2\xi}(l)=\mu F_{1\xi}(l),\quad H_1(l)=H_2(l),\quad H_{2\xi}=\mu H_{1\xi},\\ A_{2\xi}(l)-kA_{1\xi}(l)=2l^{-2}\mathrm{E}A_1(l)F_1(l),\quad C_{2\xi}(l)-kC_{1\xi}(l)=2l^{-2}\mathrm{E}C_1(l)F_1(l),\end{gathered} \tag{41}$$

where

$$\begin{gathered}F_1(l)=\frac{1}{\mathrm{P}_1}\left[-\frac{1}{6}N_{11}+\mathrm{G}_1\left(\frac{(\alpha_1+\gamma_1)l^2}{40}+\frac{(A_1^s+C_1^s)l}{12}\right)\right],\\ A_1(l)=\alpha_1 l+A_1^s,\quad C_1(l)=\gamma_1 l+C_1^s.\end{gathered} \tag{42}$$

Further,

$$\begin{gathered}F_2(l)=-\frac{\chi(l-1)^2}{6\mathrm{P}_2l^2}\left[\frac{\mathrm{G}_2(\alpha_2+\gamma_2)(l-1)}{2}+N_{21}\right],\\ H_1(l)=\frac{1}{2\mathrm{P}_1}\left[-\frac{\mathrm{G}_1(\alpha_1-\gamma_1)l^2}{20}+\frac{\mathrm{G}_1(A_1^s-C_1^s)l}{6}-\frac{N_{12}}{3}\right],\\ H_2(l)=-\frac{\chi(l-1)^2}{6\mathrm{P}_2l^2}\left[\frac{\mathrm{G}_2(\alpha_2-\gamma_2)(l-1)}{2}+N_{22}\right],\\ F_{1\xi}(l)=\frac{1}{\mathrm{P}_1l}\left[-\frac{2}{3}N_{11}+\frac{3}{20}\mathrm{G}_1(\alpha_1+\gamma_1)l^2+\frac{5\mathrm{G}_1}{12}(A_1^s+C_1^s)l\right],\\ F_{2\xi}(l)=-\frac{\chi(l-1)}{\mathrm{P}_2l^2}\left[\frac{2}{3}N_{21}+\frac{\mathrm{G}_2(\alpha_2+\gamma_2)(l-1)}{4}\right],\\ H_{1\xi}(l)=\frac{1}{\mathrm{P}_1l}\left[-\frac{2}{3}N_{12}+\frac{3}{20}\mathrm{G}_1(\alpha_1-\gamma_1)l^2+\frac{5\mathrm{G}_1}{12}(A_1^s-C_1^s)l\right],\\ H_{2\xi}(l)=-\frac{\chi(l-1)}{\mathrm{P}_2l^2}\left[\frac{2}{3}N_{22}+\frac{\mathrm{G}_2(\alpha_2-\gamma_2)(l-1)}{4}\right].\end{gathered} \tag{43}$$

Now from the first two equalities (41) we find N_{11} and N_{21}; from the last two equalities (41) we find N_{12} and N_{22}; from the last two equalities, taking into account the formulas (40), we define $\alpha_1+\gamma_1$, $\alpha_1-\gamma_1$, and therefore α_1, γ_1. Below we find the indicated values for $A_1^s=C_1^s$. This is the case of radial heating of the substrate. Here $\alpha_2+\gamma_2=(\alpha_1+\gamma_1)l+2A_1^s$, $\alpha_2-\gamma_2=(\alpha_1-\gamma_1)l$. Let's consider the simplest option: $\alpha_1=\gamma_1$ $(A_1(\xi)=C_1(\xi))$. Then $\alpha_2+\gamma_2=2(\alpha_1+A_1^s)$, $\alpha_2=\gamma_2$ and the formulas (37)–(43) are greatly simplified. Unknown will be α_1, N_{11}, N_{21}, N_{12}, N_{22}. Calculations show that in the general case

$$N_{12}=N_{22}=0,\quad N_{11}=K_1\alpha_1+K_2A_1^s,\quad N_{21}=K_3\alpha_1+K_4A_1^s, \tag{44}$$

where

$$\begin{gathered}K_1=\frac{\mathrm{G}_1l^2}{20}-\frac{1}{6(l+\mu(1-l))}\left[\frac{3\mathrm{G}_1l^3}{10}\left(1+\frac{3\mu}{2l}(1-l)\right)+\frac{\mathrm{G}_2\nu}{4}(l-1)^3\right],\\ K_2=\frac{\mathrm{G}_1l}{6}-\frac{1}{6(l+\mu(1-l))}\left[\mathrm{G}_1l^2\left(1+\frac{5\mu}{4l}(1-l)\right)+\frac{\mathrm{G}_2\nu}{4l}(l-1)^3\right],\\ K_3=\frac{\rho l^2}{(1-l)(l+\mu(1-l))}\left[\frac{3\mathrm{G}_1l^2}{20}+\frac{\mathrm{G}_2(l-1)^2}{\rho l}\left(\frac{3l}{4}+\mu(1-l)\right)\right],\\ K_4=\frac{\rho l^2}{(1-l)(l+\mu(1-l))}\left[\frac{\mathrm{G}_1l}{4}+\frac{\mathrm{G}_2(l-1)^2}{\rho l}\left(\frac{3l}{4}+\mu(1-l)\right)\right],\quad \rho=\frac{\rho_1}{\rho_2}.\end{gathered} \tag{45}$$

The constant α_1 is the solution of the quadratic equation

$$\mathrm{E}K_1\alpha_1^2 + \left(\frac{k\mathrm{P}_1 l}{2} + \mathrm{E}A_1^s(K_2 + K_1 l^{-1})\right)\alpha_1 + \mathrm{E}K_2 A_1^s l^{-1} = 0. \tag{46}$$

If the quantity

$$\delta = \left(\frac{k\mathrm{P}_1 l}{2} + \mathrm{E}A_1^s(K_2 + K_1 l^{-1})\right)^2 - 4\mathrm{E}^2 K_1 K_2 A_1^s l^{-1} \tag{47}$$

is positive, then there are two solutions of the equation (46), which means that there are two stationary solutions to the two-layer system. For $\delta = 0$ there is one stationary solution, and for $\delta < 0$ there are no solutions.

Remark 4. Fpr $l = \mu(1+\mu)^{-1}$ we get $N_{22} = 1$, $N_{12} = (1-l)(\rho l)^{-1}$, and the formulas (44), (45) retain their form with the replacement of μ by $\mu = l(1-l)^{-1}$.

As for the functions $Q_j(\xi)$, they are determined by the formulas

$$\begin{gathered} Q_1(\xi) = Q_1^s + a\xi - \frac{2}{l^2}\left(\frac{\alpha_1\xi^3}{3} + A_1^s\xi^2\right), \quad 0 \leqslant \xi \leqslant l, \\ Q_2(\xi) = b + \frac{2}{l^2}(\alpha_1 l + A_1^s)(2\xi - \xi^2), \quad l \leqslant \xi \leqslant 1, \end{gathered} \tag{48}$$

where

$$\begin{gathered} a = \frac{4}{l^2}\left[k + \frac{2\mathrm{E}F_1(l)}{l^2}\right]^{-1}\left[(k + l - l^2)(\alpha_1 l + A_1^s) + \frac{2\mathrm{E}F_1(l)}{l}\left(\frac{\alpha_1 l}{2} + A_1^s\right)\right], \\ b = Q_1^s + a_1 l + \frac{2}{3}(l-6)\alpha_1 - 2A_1^s(l+2)l^{-1}. \end{gathered} \tag{49}$$

In (48), (49) Q_1^s is the dimensionless temperature on the substrate at the origin of coordinates, and $F_1(l)$ is given by the equality (42) at $\alpha_1 = \gamma_1$, $\alpha_2 = \gamma_2 = \alpha_1 l + A_1^s$, $A_1^s = C_1^s$, and α_1 is a solution to the equation (46).

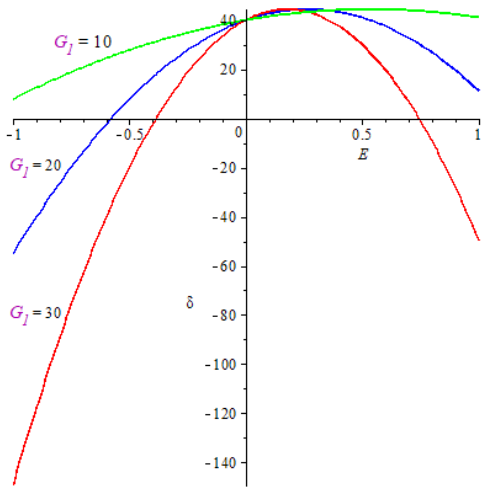


Fig. 1. Dependence $\delta(E)$ for various Grashof numbers G_1; $A_1^s = 0.1$

Figs. 1–3 shows the dependences $\delta(E)$ for various values of dimensionless parameters. All calculations are given for the transformer oil – formic acid system. The dimensionless parameters

of the physical system are as follows: $\rho = 0.74$, $\nu = 15.41$, $\chi = 0.71$, $k = 0.41$, $\beta = \beta_1/\beta_1^{-1} = 1.46$, $\mathrm{P}_1 = 308.2$, $\mathrm{P}_2 = 14.2$. Fig. 1 represent the dependence $\delta(E)$ or various Grashof numbers G_1, $\mathrm{G}_2 = \beta \mathrm{G}_1$. It can be seen that as G_1 grows, the region of existence of two solutions decreases.

Fig. 2 illustrates the dependence $\delta(E)$ for various values of the dimensionless parameter A_1^s. Here, for certain values of the parameter E, as A_1^s grows, the region where there are no solutions increases. In the case when $A_1^s \leqslant 0$ there are always two solutions. Fig. 3 shows the dependence $\delta(E)$ from the geometric parameter $l = l_1 l_2^{-1} < 1$. In this case, with an increase in the thickness of the lower layer, the region of existence of two solutions increases.

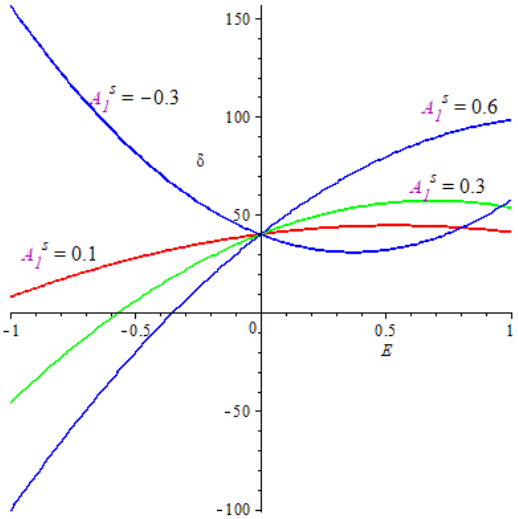


Fig. 2. Dependence $\delta(E)$ for various parameter values A_1^s

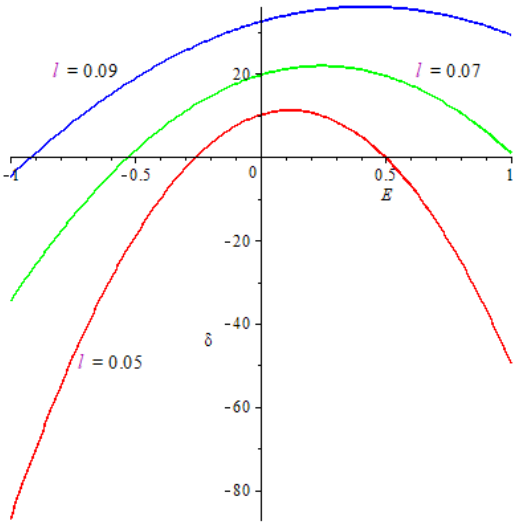


Fig. 3. Dependence $\delta(E)$ for various values of the geometric parameter l

Conclusion

In the article, the problem of three-dimensional two-layer motion with a special velocity field is reduced to the inverse conjugate problem for a system of one-dimensional integro-differential equations. In the case of a stationary flow at low Marangoni numbers, the solution is obtained in the analytical form. It is shown that, depending on the physical and geometric parameters, two stationary modes can exist. For the transformer oil - formic acid system, the effect of changes in interfacial internal energy on the number of stationary solutions has been studied.

This research was supported by the Russian Foundation for Basic Research (20-01-00234) and Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation in the framework of the establishment and development of regional Centers for Mathematics Research and Education (Agreement no. 075-02-2020-1631).

References

[1] V.K.Andreev, Yu.A.Gaponenko, O.N.Goncharova, V.V.Pukhnachev, Mathematical Models of Convection, Berlin, Boston, De Gruyter, 2020.

[2] N.Aristov, D.V.Knyazev, A.D.Polyanin, Exact solutions of the Navier-Stokes equations with the linear dependence of velocity components on two space variables, *Theoretical Foundations of Chemical Engineering*, **43**(2009), no. 5, 642–662. DOI: 10.1134/S0040579509050066

[3] V.K.Andreev, V.E.Zahvataev, E.A.Ryabitskii, Thermocapillary Instability, Nauka, Siberian brunch, Novosibirsk, 2000 (in Russian).

Об одном ползущем трехмерном конвективном движении жидкостей с изотермической границей раздела

Виктор К. Андреев
Институт вычислительного моделирования СО РАН
Красноярск, Российская Федерация

Аннотация. В работе рассматривается двухслойное трехмерное движение жидкостей, поле скоростей которых имеет специальный вид. Возникающая сопряжённая начально-краевая задача для модели Обербека–Буссинеска сведена к системе десяти интегродифференциальных уравнений с полными условиями на плоской поверхности раздела. Показано, что для малых чисел Марангони её стационарный аналог может иметь до двух решений, которые находятся в явном виде. Отдельно проанализирован случай, когда стационарное течение возникает за счет изменения внутренней межфазной энергии.

Ключевые слова: модель Обербека-Буссинеска, межфазная энергия, ползущее течение, обратная задача.

DOI: 10.17516/1997-1397-2020-13-6-670-677
УДК 532.542

Three-Dimensional Simulation of a Tank Filling With a Viscous Fluid Using the VOF Method

Evgeny I. Borzenko*
Efim I. Hegaj†
Tomsk State University
Tomsk, Russian Federation

Received 13.06.2020, received in revised form 04.09.2020, accepted 04.10.2020

Abstract. This paper presents the results of 3D modeling of a Newtonian fluid flow with a free surface. The PLIC-VOF algorithm, which is developed to solve the problems of two-dimensional fluid flows with a free surface, is generalized to the case of three-dimensional flows. Efficiency of the developed algorithm and reliability of the obtained results are justified by comparing with available data in literature and by testing approximation convergence.

Parametric calculations of a rectangular channel filling show that the free surface assumes a steady convex shape over time and then moves along the channel at a constant velocity. As a result of parametric studies, the dependences of geometric characteristics of the free surface shape on problem parameters have been plotted.

Keywords: Newtonian fluid flow, filling of a rectangular channel, free surface, 3D modeling, numerical simulation, VOF method, flow structure.

Citation: E.I. Borzenko, E.I. Hegaj, Three-Dimensional Simulation of a Tank Filling With a Viscous Fluid Using the VOF Method, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 670–677.
DOI: 10.17516/1997-1397-2020-13-6-670-677.

Introduction

Technological processes associated with casting of items, filling of tanks or draining of polymer compounds are characterized by the presence of a free surface in a fluid flow. Adequate engineering for such processes requires a detailed study of the flow features and free surface behavior [1].

Tracking of free surface evolution in a fluid flow is known as a complex problem in hydromechanics [2]. One of the first successful attempts to determine free surface dynamics in the two-dimensional approximation has been made in [3]. In this paper, a numerical method based on the VOF (Volume of Fluid) approach is proposed, which allows one to determine a free surface position at any time instant using a scalar function defined in the cells of a regular grid. Moreover, at a discrete level, the free boundary in a control volume is represented as a segment, which is parallel to one of the volume's faces. The method was further developed in [4]. A modification of the method, namely PLIC-VOF (Piecewise-Linear Interface Calculation), is proposed, according to which a free surface in a control volume is represented as a set of arbitrary oriented segments. The PLIC-VOF method has been successfully applied for two-dimensional flows in [5–10].

Many fluid flow features cannot be taken into account nor adequately assessed when modeling the process in the plane or axisymmetric approximation. Two-dimensional problem formulations

*borzenko@ftf.tsu.ru https://orcid.org/0000-0001-6264-1776
†efim_h@ftf.tsu.ru https://orcid.org/0000-0002-7973-9435

allow one to study flow mechanics only for some channel designs. However, most of the real pipelines have complex three-dimensional geometry. Therefore, there is a need to develop algorithms for calculating and modeling fluid flows in a full three-dimensional formulation.

In this work, testing of a modified VOF method on the three-dimensional fluid flow modeling is implemented.

1. Formulation of the problem

A three-dimensional flow, which occurs when a vertical rectangular channel is being filled in a gravity field, is considered. A solution domain is schematically shown in Fig. 1 *a*. In this case, the fluid is supplied from the bottom through the inlet section at given constant flow rate.

At the initial time instant, the channel is partially filled with a fluid, and the free surface is represented as a plane $z = const$ confined by the walls.

When filling the channel, the free surface becomes curved and assumes a convex shape. The maximum height of the free surface, χ (Fig. 1 *b*), is taken as convexity characteristics, which is determined by the values of the parameters Re and W.

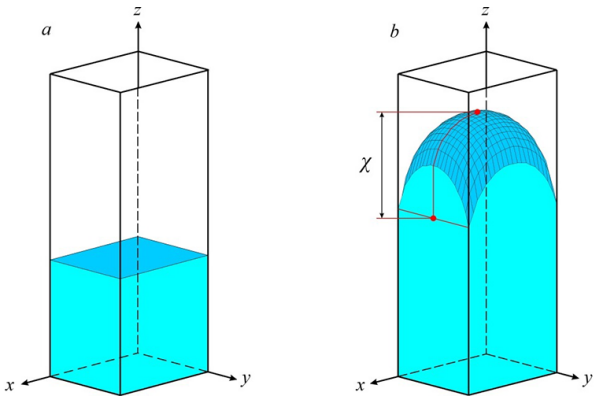


Fig. 1. Solution domain (a) at the initial time instant and (b) during the filling process

Mathematical formulation of the problem includes the Navier-Stokes and continuity equations written in a dimensionless form as

$$\begin{cases} \mathrm{Re}\left(\dfrac{dU_x}{dt} + U_x\dfrac{dU_x}{dx} + U_y\dfrac{dU_x}{dy} + U_z\dfrac{dU_x}{dz}\right) = -\dfrac{dP}{dx} + \left(\dfrac{d^2U_x}{dx^2} + \dfrac{d^2U_x}{dy^2} + \dfrac{d^2U_x}{dz^2}\right) \\ \mathrm{Re}\left(\dfrac{dU_y}{dt} + U_x\dfrac{dU_y}{dx} + U_y\dfrac{dU_y}{dy} + U_z\dfrac{dU_y}{dz}\right) = -\dfrac{dP}{dy} + \left(\dfrac{d^2U_y}{dx^2} + \dfrac{d^2U_y}{dy^2} + \dfrac{d^2U_y}{dz^2}\right) \\ \mathrm{Re}\left(\dfrac{dU_z}{dt} + U_x\dfrac{dU_z}{dx} + U_y\dfrac{dU_z}{dy} + U_z\dfrac{dU_z}{dz}\right) = -\dfrac{dP}{dz} + \left(\dfrac{d^2U_z}{dx^2} + \dfrac{d^2U_z}{dy^2} + \dfrac{d^2U_z}{dz^2}\right) - \mathrm{W} \end{cases} \tag{1}$$

$$\frac{dU_x}{dx} + \frac{dU_y}{dy} + \frac{dU_z}{dz} = 0 \tag{2}$$

No-slip conditions are assigned on the solid walls. On the free surface, the continuity conditions for normal and shear stresses are used. The fluid is supplied through the inlet section at a velocity equal to unity.

The following quantities are used as length, velocity, time, and pressure scales: L (a characteristic size of the channel), U_0 (an average velocity at the inlet section), and the complexes of L/U_0 and $\mu U_0/L$, respectively. The problem formulation includes dimensionless criteria: the

Reynolds number $\mathrm{Re} = \rho U_0 L/\mu$ and the parameter $\mathrm{W} = \rho L^2 g/\mu U_0 = \mathrm{Re}/\mathrm{Fr}$, which is equal to the ratio of the Reynolds and Froude numbers.

2. Method of solving

The formulated problem is solved numerically using the finite volume method implemented on a staggered grid. Kinematic and dynamic characteristics of the flow are determined using the SIMPLE algorithm [11]. In this case, to approximate of convective and non-stationary term, an exponential scheme was used. Tracking of the free surface evolution is carried out by a modified VOF method, which is generalized to the three-dimensional case with account for an arbitrary inclination of the free surface in a control volume.

The VOF method implies introducing of a scalar function F, whose value is equal to unity at all points occupied by the fluid and equal to zero at the rest of the points. At a discrete level, when averaging over the control volume, the value of F is equal to a volume fraction of the control volume occupied by the fluid. In particular, when the control volume is entirely filled with a fluid, $F = 1$, and when the control volume does not contain any fluid, $F = 0$. If there is a free surface in the control volume, $0 < F < 1$. The values of this function can be determined from the equality to zero of the total derivative of F with respect to time, which reflects the law of conservation of mass

$$\frac{dF}{dt} + U_x \frac{dF}{dx} + U_y \frac{dF}{dy} + U_z \frac{dF}{dz} = 0 \tag{3}$$

When integrating this equation over the control volume, it is necessary to determine the fluxes through the faces, which are calculated using the obtained velocity values on these faces and the orientation of the free surface in the control volume.

It is assumed that at a discrete level, the free surface in the control volume represents a cutting plane, whose position is determined by its normal and by the fraction of the volume filled with the fluid. In the three-dimensional case, there are eight options for the free surface to cross the control volume (Fig. 2), which are characterized by positive components of the normal vector. The normal to the free surface is assumed to be directed along the gradient of the function F. When the value of F and the direction of the normal to the free surface are known, the plane approximating the free surface can be drawn in the boundary control volume. Therefore, in addition to the boundary control volume tracking, the function F is used to detect the fluid location inside the volume.

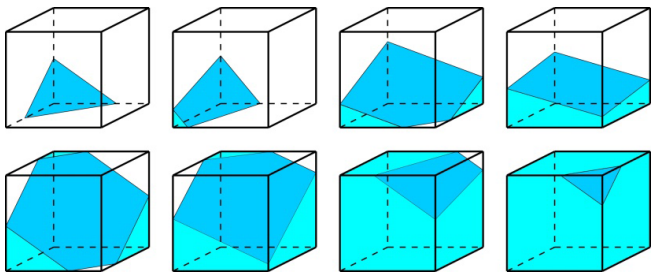


Fig. 2. Possible locations for the free surface inside the control volume

Other possible free surface locations are reduced to those shown in Fig. 2 by rotating the control volume and / or by reflecting in coordinate planes.

Variation of the function F with time is determined by equation (3), which can be solved numerically after calculating its fluxes through the control volume faces. An illustration of the method in use is shown in Fig. 3. Two adjacent control volumes are considered, where the fluid

flows through a common side. The velocity U on the adjacent face determines which one is a donor and which is an acceptor. Afterwards, a plane, which is parallel to the common side, is plotted at a distance of $U\Delta t$. The fluid fraction in the donor cell ΔF, which is enclosed between the plotted plane and the adjacent face, is transferred to the acceptor. The value of this fluid fraction (ΔF) is calculated by analytical formulas for polyhedra volume.

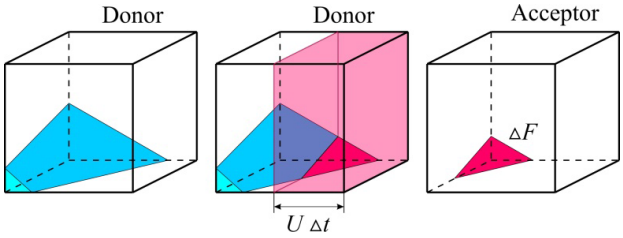


Fig. 3. Illustration of the VOF method operation

The fluxes through the other five faces of the control volume are determined similarly.

3. Verification of numerical method

To verify operational capability of the developed algorithm and reliability of the obtained results, the approximation convergence is tested on a sequence of grids.

The calculations show that at a time instant of $t = 2$ in a cross section of $z = 2$, the absolute value of the maximum transverse velocity is of the order of 10^{-6}. It is supposed that the inlet boundary and the free surface do not affect the flow in this section, where a steady-state flow is observed. Thus, to verify the obtained velocity distributions, a well-known solution is used [12]:

$$\widetilde{U}_z = 3.665 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^3} \left[1 - \frac{\operatorname{ch} \dfrac{(2n+1)\,\pi\,(2y-1)}{2}}{\operatorname{ch} \dfrac{(2n+1)\,\pi}{2}} \right] \cos \frac{(2n+1)\,\pi\,(2x-1)}{2}. \tag{4}$$

The velocity error can be calculated as

$$\Delta U_z = \max \left| \frac{\widetilde{U}_z - U_{z=2}}{\widetilde{U}_z} \right| 100\%. \tag{5}$$

Since the fluid velocity is equal to unity in the inlet section, at a time instant of $t = 2$, the volume of the inflowed fluid should equal $V = 2$. The error in the calculation of the volume is determined by formula

$$\Delta V = \max \left| \frac{2 - V_{t=2}}{2} \right| 100\%. \tag{6}$$

To show the approximation convergence, a velocity profile along a straight line of $y = 0.5$ at $z = 2$ (Fig. 4 *a*) and a free surface shape in a cross section of $y = 0.5$ (Fig. 4 *b*) are calculated at different grid steps.

According to Fig. 4 *b*, a maximal difference in the free surface shapes is observed on the solid walls. Therefore, to control the approximation convergence, a parameter H (the free surface height at the point of $x = 0,\ y = 0.5$) is introduced.

Thus, the errors in the calculations of the fluid velocity, fluid volume, and free surface height on the solid wall H are selected as controlled characteristics in the computational method verification.

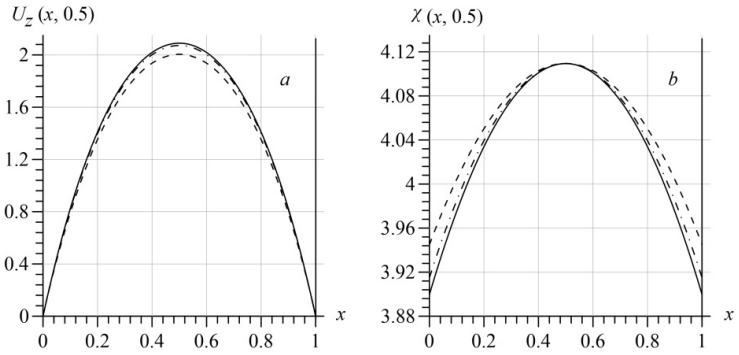


Fig. 4. (a) Velocity distribution along a straight line of $y = 0.5$ at $z = 2$ and (b) the free surface shape in a cross section of $y = 0.5$ at a time instant of $t = 2$ for Re = 0.1 and W = 32: $h = 0.1$ (the dashed line), $h = 0.05$ (the dotted and dashed line), and $h = 0.025$ (the solid line)

The obtained results presented in Tab. 1 demonstrate the approximation convergence for the selected characteristics.

Table 1. Dependence of the values of the controlled characteristics on the grid step at $t = 2$, Re = 0.1, W = 32

h	$\Delta U_z, \%$	$\Delta V, \%$	H
0.1	2.53	0.42	3.944
0.05	2.15	0.41	3.915
0.025	2.01	0.40	3.900

4. Results of calculations

The initially flat free surface becomes curved over time and assumes a convex shape, which then moves upward along the channel and remains unchanged. Fig. 5 shows the evolution of the free surface shape.

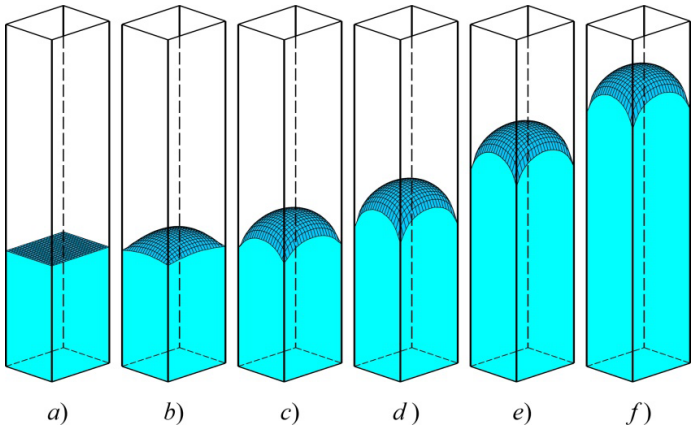


Fig. 5. Free surface shapes for Re = 0.1, W = 32 at various time instants: (a) $t = 0$, (b) $t = 0.2$, (c) $t = 0.5$, (d) $t = 1$, (e) $t = 2$, and (f) $t = 3$

Fig. 6 demonstrates distributions of the velocities and pressure in a longitudinal section of the channel $y = 0.5$ for Re = 0.1 and W = 32 at a time instant of $t = 2$. A fountain flow is observed in the section under consideration. Three zones can be distinguished in the flow: a hydrodynamic flow stabilization zone near the inlet section; a fountain flow zone near the free surface; and a one-dimensional flow zone. The calculations showed that the lengths of the stabilization and fountain flow regions are less than unity for the selected values of the dimensionless criteria.

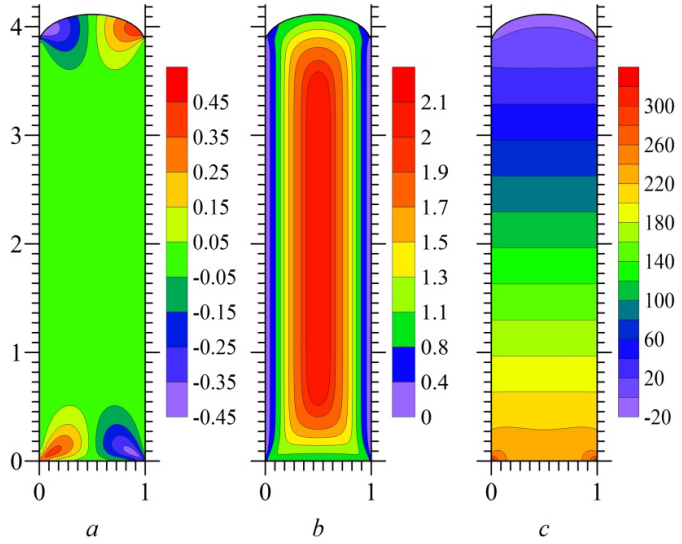


Fig. 6. Distribution of the kinematic characteristics along the channel for Re = 0.1, W = 32 at a time instant of $t = 2$ in a cross section of $y = 0.5$: (a) velocity U_x, (b) velocity U_z, and (c) pressure

In the other longitudinal sections, distributions of the kinematic characteristics qualitatively coincide with those presented in Fig. 6. In the section of $x = y$, velocity distributions (U_x and U_y) coincide with each other, which also confirms the efficiency of the computational algorithm.

The obtained results are compared with those of other authors. In particular, the results of studying of a channel filling in the creeping flow approximation are presented in [12]. Fig. 7 a shows the calculated values of the free surface height as a function of the parameter W presented in [13] (the solid line) and the results obtained by the VOF method (the dots).

Fig. 7 b illustrates a difference in the free surface convexity depending on the parameter W.

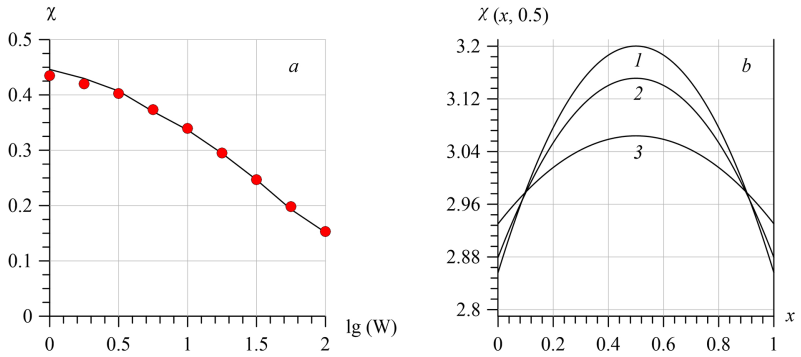


Fig. 7. (a) Comparison of the dependence of free surface convexity on the parameter W in the current work (the dots) and in [13] (the solid line); (b) free surface shapes in a cross section of $y = 0.5$ for Re = 0.1 at the same time instant: *1* — W = 0, *2* — W = 10, and *3* – W = 100

The maximal difference in the calculated results does not exceed 3%. As a result of comparison, qualitative agreement and little quantitative deviations are observed.

Conclusion

In this work, testing of a modified VOF method on the three-dimensional fluid flow modeling is implemented. In particular, the PLIC-VOF algorithm, which is developed to solve the problems of two-dimensional fluid flows with a free surface, is generalized to the case of three-dimensional flows. Comparing with available data in literature and by testing approximation convergence are justified by efficiency of the developed algorithm. As a demonstration of the operation of the calculation program, the results of a study on filling a rectangular channel are given. Parametric calculations show that the free surface assumes a steady convex shape over time and then moves along the channel at a constant velocity.

The research is implemented at the expanses of the Russian Science Foundation (project no. 18-19-00021).

References

[1] I.A.Glushkov et al., Simulation of the molding of articles from free-molding compositions, Moscow, Arkhitektura-S, 2007 (in Russian).

[2] N.D.Katopodes, Free-Surface Flow: Computational Methods, 2019.

[3] B.D.Nichols, C.W.Hirt, R.S.Hotchkiss, SOLA-VOF: a solution algorithm for transient fluid flow with multiple free boundaries, Los Alamos Scientific Laboratory Report, LA-8355, 1980.

[4] W.Jang, J.Jilesen, F.S.Lien, H.Ji, A study on the extension of a VOF/PLIC based method to a curvilinear co-ordinate system, *International Journal of Computational Fluid Dynamics*, **22**(2008), no. 4, 241–257.

[5] A.Issakhov, Y.Zhandaulet, A.Nogaeva, Numerical simulation of dam break flow for various forms of the obstacle by VOF method, *International Journal of Multiphase Flow*, **109**(2018), 191–206. DOI: 10.1016/J.IJMULTIPHASEFLOW.2018.08.003

[6] D.M.Hargreaves, H.P.Morvan, N.G.Wright, Validation of the Volume of Fluid Method for Free Surface Calculation: The Broad-Crested Weir, *Engineering Applications of Computational Fluid Mechanics*, **1**(2007), no. 2, 136–146. DOI: 10.1080/19942060.2007.11015188

[7] S.Hansch, D.Lucas, T.Hohne, E.Krepper, Gu.Montoya, Comparative Simulations of Free Surface Flows Using VOF-Methods and a New Approach for Multi-Scale Interfacial Structures, Proceedings of the ASME 2013 Fluids Engineering Summer Meeting (FEDSM2013-16104), 2013. DOI: 10.1115/FEDSM2013-16104

[8] S.Saincher, J.Banerjee, A Redistribution-Based Volume Preserving PLIC-VOF Technique, *Numerical Heat Transfer*, **67**(2015), 338–362. DOI: 10.1080/10407790.2014.950078

[9] X.Yin, I.Zaricos, N.K.Karadimitriou, A.Raoof, S.M.Hassanizadeh, Direct simulations of two-phase flow experiments of different geometry complexities using Volume-of-Fluid (VOF) method, *Chemical Engineering Science*, **195**(2018) 820–827.
DOI: 10.1016/j.ces.2018.10.029.

[10] M.Karimi, H.Droghetti, D.L.Marchisio, Multiscale Modeling of Expanding Polyurethane Foams via Computational Fluid Dynamics and Population Balance Equation, *Macromol. Symp.*, **360**(2016), 108–122. DOI: 10.1002/masy.201500108

[11] L.G.Loytsyansky, Fluid and Gas Mechanics, Moscow-Leningrad, Gostekhizdat, 1950.

[12] S.Patankar, Numerical heat transfer and fluid flow, Hemisphere Publishing Corporation, 1980.

[13] G.R.Shrager, A.N.Kozlobrodov, V.A.Yakutenok, Modeling of hydrodynamic processes in polymer processing technology, Tomsk, TGU Publ., 1999.

Моделирование пространственного заполнения емкости вязкой жидкостью с использованием VOF-метода

Евгений И. Борзенко
Ефим И. Хегай
Томский государственный университет
Томск, Российская Федерация

Аннотация. В настоящей работе представлены результаты моделирования пространственного течения ньютоновской жидкости со свободной поверхностью. Алгоритм PLIC VOF, предназначенный для решения задач о течении жидкостей со свободной поверхностью в двумерной постановке, обобщен на случай пространственных потоков. Работоспособность разработанного алгоритма и достоверность получаемых результатов продемонстрированы путем сравнения с литературными данными и проверкой аппроксимационной сходимости.

Параметрические расчеты заполнения канала с прямоугольным сечением показали, что с течением времени свободная поверхность принимает установившуюся выпуклую форму, которая перемещается вдоль канала с постоянной скоростью. В результате параметрического исследования построены зависимости геометрических характеристик формы свободной поверхности от параметров задачи.

Ключевые слова: течение ньютоновской жидкости, заполнение прямоугольного канала, свободная поверхность, трехмерное моделирование, численное моделирование, VOF-метод, структура потока.

DOI: 10.17516/1997-1397-2020-13-6-678-693
УДК 517.9

On Asymptotic Dynamical Regimes of Manakov N-soliton Trains in Adiabatic Approximation

Vladimir S. Gerdjikov*
National Research Nuclear University MEPHI
Moscow, Russian Federation
Institute of Mathematics and Informatics Bulgarian Academy of Sciences
Sofia, Bulgaria
Institute for Advanced Physical Studies, New Bulgarian University
Sofia, Bulgaria

Michail D. Todorov†
San Diego State University
San Diego, CA, USA
Technical University of Sofia
Sofia, Bulgaria

Received 10.06.2020, received in revised form 14.08.2020, accepted 20.09.2020

Abstract. We analyze the dynamical behavior of the N-soliton train in the adiabatic approximation of the Manakov model. The evolution of Manakov N-soliton trains is described by the complex Toda chain (CTC) which is a completely integrable dynamical model. Calculating the eigenvalues of its Lax matrix allows us to determine the asymptotic velocity of each soliton. So we describe sets of soliton parameters that ensure one of the two main types of asymptotic regimes: the bound state regime (BSR) and the free asymptotic regime (FAR). In particular we find explicit description of special symmetric configurations of N solitons that ensure BSR and FAR. We find excellent matches between the trajectories of the solitons predicted by CTC with the ones calculated numerically from the Manakov system for wide classes of soliton parameters. This confirms the validity of our model.

Keywords: Manakov model, soliton interactions, adiabatic approximations complex Toda chain

Citation: V.S. Gerdjikov, M.D. Todorov, On Asymptotic Dynamical Regimes of Manakov **N**-soliton Trains in Adiabatic Approximation, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 678–693.
DOI: 10.17516/1997-1397-2020-13-6-678-693.

1. Introduction and preliminaries

The solitons and their interactions find numerous applications of in many areas of today nonlinear physics, such as hydrodynamics, nonlinear optics, Bose-Einstein condensates, etc. [1–3, 7, 27, 28, 34]. This explains why it is important to study their interactions. The first results on soliton interactions were obtained by Zakharov and Shabat [35, 36]. There they proved that the nonlinear Schrödinger equation

$$iu_t + \frac{1}{2}u_{xx} + |u|^2 u(x,t) = 0. \tag{1}$$

*vgerdjikov@math.bas.bg https://orcid.org/0000-0002-1058-6565
†mtod@tu-sofia.bg https://orcid.org/0000-0002-4019-5880

can be integrated by the inverse scattering method (ISM). Then they constructed the N-soliton solution of (1) and calculated their limits for $t \to \infty$ and $t \to -\infty$, assuming that all solitons have different velocities. Comparing the asymptotics they concluded that the soliton interactions are purely elastic, i.e., no new solitons can be created. In addition the solitons preserve their amplitudes and velocities, and the only effect of the interactions are relative shifts of the center of masses and phases.

Later Karpman and Solov'ev proposed another approach to the soliton interactions based on the adiabatic approximation [25, 26]. They proposed to model the N-soliton trains of the NLS eq. (1). By N-soliton train they meant a solution of the NLS eq. with initial condition:

$$\begin{aligned} u(x,t=0) &= \sum_{k=1}^{N} \vec{u}_k(x,t=0), & u_k(x,t) &= \frac{2\nu_k e^{i\phi_k}}{\cosh(z_k)}, \\ z_k &= 2\nu_k(x-\xi_k(t)), & \xi_k(t) &= 2\mu_k t + \xi_{k,0}, \\ \phi_k &= \frac{\mu_k}{\nu_k} z_k + \delta_k(t), & \delta_k(t) &= 2(\mu_k^2+\nu_k^2)t + \delta_{k,0}. \end{aligned} \tag{2}$$

The adiabatic approximation holds true if the soliton parameters satisfy [26]:

$$|\nu_k - \nu_0| \ll \nu_0, \quad |\mu_k - \mu_0| \ll \mu_0, \quad |\nu_k - \nu_0||\xi_{k+1,0} - \xi_{k,0}| \gg 1, \tag{3}$$

where $\nu_0 = \frac{1}{N}\sum\limits_{k=1}^{N} \nu_k$, and $\mu_0 = \frac{1}{N}\sum\limits_{k=1}^{N} \mu_k$ are the average amplitude and velocity respectively. In fact we have two different scales:

$$|\nu_k - \nu_0| \simeq \varepsilon_0^{1/2}, \qquad |\mu_k - \mu_0| \simeq \varepsilon_0^{1/2}, \qquad |\xi_{k+1,0} - \xi_{k,0}| \simeq \varepsilon_0^{-1}.$$

In this approximation the dynamics of the N-soliton train is described by a dynamical system for the $4N$ soliton parameters. What Karpman and Solov'ev did was to derive the dynamical system for the two soliton interactions: a system of 8 equations for the 8 soliton parameters. They were able also to solve it analytically.

Later their results were generalized to N-soliton trains [16, 17, 24]. The corresponding model can be written down in the form :

$$\begin{aligned} \frac{d\lambda_k}{dt} &= -4\nu_0 \left(e^{Q_{k+1}-Q_k} - e^{Q_k - Q_{k-1}} \right), \\ \frac{dQ_k}{dt} &= -4\nu_0 \lambda_k, \end{aligned} \tag{4}$$

where $\lambda_k = \mu_k + i\nu_k$ and

$$\begin{aligned} Q_k &= -2\nu_0\xi_k + k\ln 4\nu_0^2 - i(\delta_k + \delta_0 + k\pi - 2\mu_0\xi_k), \\ \nu_0 &= \frac{1}{N}\sum_{s=1}^{N}\nu_s, \qquad \mu_0 = \frac{1}{N}\sum_{s=1}^{N}\mu_s, \qquad \delta_0 = \frac{1}{N}\sum_{s=1}^{N}\delta_s. \end{aligned} \tag{5}$$

Obviously the system (4) becomes the Toda chain with free ends for the complex variables Q_k:

$$\begin{aligned} &\frac{d^2Q_k}{dt^2} = -4\nu_0\frac{d\lambda_k}{dt} = 16\nu_0^2\left(e^{Q_{k+1}-Q_k} - e^{Q_k-Q_{k-1}}\right) \qquad k = 2,\dots,N-1, \\ &\frac{d^2Q_1}{dt^2} = 16\nu_0^2 e^{Q_2-Q_1}, \qquad \frac{d^2Q_N}{dt^2} = -16\nu_0^2 e^{Q_N-Q_{N-1}}. \end{aligned} \tag{6}$$

which is known as the complex Toda chain (CTC).

It is well known that the standard (real) Toda chain is an integrable system [9, 28, 31]. In the case of (6), which is known as Toda chain with open ends, it was possible to write down its solutions explicitly [31]. An important fact is that these solutions depend analytically on their parameters and can be easily generalized to the CTC.

In fact some time ago a special configurations of soliton trains that are modeled by the real Toda chain [4, 5] were found. For them we must choose solitons with equal amplitudes (i.e., $\nu_k = \nu_0$), vanishing initial velocities ($\mu_k = 0$), and out-of phase $\delta_{k+1} - \delta_k = \pi$. It is easy to see that under these assumptions Q_k become real valued and (6) become the standard Toda chain.

The adiabatic approach of Karpman and Solov'ev has a drawback: it is an approximate method whose precision is determined by ε_0. On the other hand it has the advantages: first, it is not limited only to solitons with different velocities, and second, it can take into account possible perturbations of the NLS [16, 17, 24].

Another important generalization of the NLS equation is known as the Manakov model [28] (vector NLS):

$$i\vec{u}_t + \frac{1}{2}\vec{u}_{xx} + (\vec{u}^\dagger, \vec{u})\vec{u}(x,t) = 0. \tag{7}$$

The corresponding vector N-soliton train is determined by the initial condition:

$$\begin{aligned} \vec{u}(x,t=0) &= \sum_{k=1}^{N} \vec{u}_k(x,t=0), & \vec{u}_k(x,t) &= \frac{2\nu_k e^{i\phi_k}}{\cosh(z_k)}\vec{n}_k, \\ z_k &= 2\nu_k(x - \xi_k(t)), & \xi_k(t) &= 2\mu_k t + \xi_{k,0}, \\ \phi_k &= \frac{\mu_k}{\nu_k} z_k + \delta_k(t), & \delta_k(t) &= 2(\mu_k^2 + \nu_k^2)t + \delta_{k,0}, \end{aligned} \tag{8}$$

where the constant polarization vector $\vec{n}_k$ is normalized by

$$\vec{n}_k = \begin{pmatrix} \cos(\theta_k)e^{i\gamma_k} \\ \sin(\theta_k)e^{-i\gamma_k} \end{pmatrix}, \qquad (\vec{n}_k^\dagger, \vec{n}_k) = 1, \qquad \sum_{s=1}^{n} \arg \vec{n}_{k;s} = 0.$$

Therefore each Manakov soliton is parametrized by 6 parameters.

It was natural to extend the Karpman-Solov'ev method to the Manakov model. The result is known as the generalized CTC (GCTC) [10–12, 14]. Of course later the GCTC was also adapted to treat the effects of several types of perturbations on solitons [8, 13, 19, 30, 32, 33].

The advantage of the integrability of the CTC and GCTC is in the fact that knowing the initial set of soliton parameters one can predict the asymptotic regime of the soliton train [16, 17, 24]. On the other hand it is possible to find the set of constraints on the soliton parameters that would ensure given asymptotic regime. These constraints were derived and analyzed for 2 and 3-soliton trains; for larger number of solitons only fragmentary results such as the quasi-equidistant propagation of solitons [16] are known.

The aim of the present paper is to reinvestigate these results and to demonstrate several configuration of multisoliton trains for which one can predict that they will go into bound state regime (BSR) or into free asymptotic regime (FAR). In Section 2 we outline the derivation of the GCTC model, see eq. (16) below which now depends also on the polarization vectors $\vec{n}_k$ and models the behavior of the N-soliton train of the vector NLS. We also formulate the Lax representation for the GCTC and explain how it can be used to determine the asymptotic regime of the soliton train. In Section 3 we formulate two classes of explicit constraints on the soliton

parameters that are responsible for BSR and FAR. The first class are generic conditions that ensure that the Lax matrix becomes either real or purely imaginary. The second class are based on special explicit constraints on the soliton parameters that make the eigenvalues of the Lax matrix proportional to each other, so it is easier to establish if they are real or purely imaginary.

2. Preliminaries

2.1. Variational approach and generalized CTC

The Lagrangian of the vector NLS perturbed by external potential is:

$$\mathcal{L}[\vec{u}] = \int_{-\infty}^{\infty} dt\ \frac{i}{2}\left[(\vec{u}\dagger, \vec{u}_t) - (\vec{u}_t^{\dagger}, \vec{u})\right] - H, \qquad H[\vec{u}] \ = \int_{-\infty}^{\infty} dx\ \left[-\frac{1}{2}(\vec{u}_x^{\dagger}, \vec{u}_x) + \frac{1}{2}(\vec{u}^{\dagger}, \vec{u})^2\right]. \quad (9)$$

Then the Lagrange equations of motion:

$$\frac{d}{dt}\frac{\delta\mathcal{L}}{\delta\vec{u}_t^{\dagger}} - \frac{\delta\mathcal{L}}{\delta\vec{u}\dagger} = 0, \tag{10}$$

coincide with the vector NLS with external potential $V(x)$.

Next we insert $\vec{u}(x,t) = \sum_{k=1}^{N} \vec{u}_k(x,t)$ (see eq. (8)) and integrate over x neglecting all terms of order ϵ and higher. In doing this we assume that $\xi_1 < \xi_2 < \cdots < \xi_N$ at $t = 0$ and use the fact, that only the nearest neighbor solitons will contribute. All integrals of the form:

$$\int_{-\infty}^{\infty} dx\ (\vec{u}_{k,x}^{\dagger}, \vec{u}_{p,x}), \qquad \int_{-\infty}^{\infty} dx\ (\vec{u}_k^{\dagger}, \vec{u}_p), \tag{11}$$

with $|p-k| \geqslant 2$ can be neglected. The same holds true also for the integrals

$$\int_{-\infty}^{\infty} dx\ (\vec{u}_k^{\dagger}, \vec{u}_p)(\vec{u}_s^{\dagger}, \vec{u}_l),$$

where at least three of the indices k, p, s, l have different values. In doing this key role play the following integrals:

$$\begin{aligned} \mathcal{J}_2(a) &= \int_{-\infty}^{\infty} \frac{dz\, e^{iaz}}{2\cosh^2 z} = \frac{\pi a}{2\sinh\frac{a\pi}{2}}, \\ K(a,\Delta) &\equiv \int_{-\infty}^{\infty} \frac{dz\, e^{iaz}}{2\cosh z\cosh(z+\Delta)} = \frac{\pi(1-e^{-ia\Delta})}{2i\sinh(\Delta)\sinh(\pi a/2)}, \end{aligned} \tag{12}$$

Thus after long calculations we obtain:

$$\begin{aligned} &\mathcal{L} = \sum_{k=1}^{N}\mathcal{L}_k + \sum_{k=1}^{N}\sum_{n=k\pm1}\widetilde{\mathcal{L}}_{k,n}, \qquad \mathcal{L}_{k,n} = 16\nu_0^3 e^{-\Delta_{k,n}}(R_{k,n} + R_{k,n}^*), \\ &R_{k,n} = e^{i(\widetilde{\delta}_n - \widetilde{\delta}_k)}(\vec{n}_k^{\dagger}\vec{n}_n), \qquad \widetilde{\delta}_k = \delta_k - 2\mu_0\xi_k, \\ &\Delta_{k,n} = 2s_{k,n}\nu_0(\xi_k - \xi_n) \gg 1, \qquad s_{k,k+1} = -1, \qquad s_{k,k-1} = 1, \end{aligned} \tag{13}$$

where

$$\mathcal{L}_k = -2i\nu_k\left((\vec{n}_{k,t}^{\dagger}, \vec{n}_k) - (\vec{n}_k^{\dagger}, \vec{n}_{k,t})\right) + 8\mu_k\nu_k\frac{d\xi_k}{dt} - 4\nu_k\frac{d\delta_k}{dt} - 8\mu_k^2\nu_k + \frac{8\nu_k^3}{3} \tag{14}$$

The equations of motion are given by:

$$\frac{d}{dt}\frac{\delta\mathcal{L}}{\delta p_{k,t}} - \frac{\delta\mathcal{L}}{\delta p_k} = 0, \tag{15}$$

where p_k stands for one of the soliton parameters: δ_k, ξ_k, μ_k, ν_k and $\vec{n}_k^\dagger$. The corresponding system is a generalization of CTC:

$$\begin{aligned} \frac{d\lambda_k}{dt} &= -4\nu_0\left(e^{Q_{k+1}-Q_k}(\vec{n}_{k+1}^\dagger,\vec{n}_k) - e^{Q_k-Q_{k-1}}(\vec{n}_k^\dagger,\vec{n}_{k-1})\right), \\ \frac{dQ_k}{dt} &= -4\nu_0\lambda_k, \qquad \frac{d\vec{n}_k}{dt} = \mathcal{O}(\epsilon), \end{aligned} \tag{16}$$

where again $\lambda_k = \mu_k + i\nu_k$ and the other variables are given by (5). Now we have additional equations describing the evolution of the polarization vectors. But note, that their evolution is slow, and in addition the products $(\vec{n}_{k+1}^\dagger,\vec{n}_k)$ multiply the exponents $e^{Q_{k+1}-Q_k}$ which are also of the order of ϵ. Since we are keeping only terms of the order of ϵ we can replace $(\vec{n}_{k+1}^\dagger,\vec{n}_k)$ by their initial values

$$(\vec{n}_{k+1}^\dagger,\vec{n}_k)\Big|_{t=0} = m_{0k}^2 e^{2i\sigma_k}, \qquad k = 1,\dots,N-1. \tag{17}$$

We will consider most general form of the polarization vectors:

$$|\vec{n}_k\rangle = \begin{pmatrix} \cos(\theta_k)e^{i\gamma_k} \\ \sin(\theta_k)e^{-i\gamma_k} \end{pmatrix},$$

$$\begin{aligned} \langle\vec{n}_{k+1}^\dagger|\vec{n}_k\rangle &= \cos(\theta_{k+1}\cos(\theta_k)e^{-i(\gamma_{k+1}-\gamma_k)} + \sin(\theta_{k+1}\sin(\theta_k)e^{i(\gamma_{k+1}-\gamma_k)} = \rho_k e^{i\sigma_k}, \\ \rho_k^2 &= \cos^2(\gamma_{k+1}-\gamma_k)\cos^2(\theta_{k+1}-\theta_k) + \sin^2(\gamma_{k+1}-\gamma_k)\cos^2(\theta_{k+1}+\theta_k). \\ \sigma_k &= -\arctan\left(\tan(\gamma_{k+1}-\gamma_k)\frac{\cos(\theta_{k+1}+\theta_k)}{\cos(\theta_{k+1}-\theta_k)}\right), \\ a_k &= \frac{\nu_0}{2}\rho_k^2\exp(-\nu_0(\xi_{k+1}-\xi_k))\exp(-i(\delta_{k+1}-\delta_k-\sigma_k+\pi)/2). \end{aligned} \tag{18}$$

In our previous papers we considered configurations for which $|\vec{n}_k\rangle$ are real, i.e., $\gamma_k = 0$. Note that the effect of the polarization vectors could be viewed as change of the distance between the solitons and between the phases.

The system (16) was derived for the Manakov system $n = 2$ by other methods in [18]. There the GCTC model was tested numerically and found to give very good agreement with the numerical solution of the Manakov model. However the tests were done only for real values of the polarization vectors, i.e., all $\gamma_k = 0$, $k = 1,\dots,N$. Below we will take into account the effect of γ_k onto the dynamical regimes of the solitons.

2.2. Asymptotic regimes: general approach

We first briefly remind the main results concerning the CTC model [14,16–18,24]. The CTC is completely integrable model; it allows Lax representation $L_t = [A.L]$, where:

$$L = \sum_{s=1}^{N}\left(b_s E_{ss} + a_s(E_{s,s+1} + E_{s+1,s})\right), \quad A = \sum_{s=1}^{N}\left(a_s(E_{s,s+1} - E_{s+1,s})\right), \tag{19}$$

where $a_s = \exp((Q_{s+1} - Q_s)/2)$, $b_s = \mu_{s,t} + i\nu_{s,t}$ and the matrices E_{ks} are determined by $(E_{ks})_{pj} = \delta_{kp}\delta_{sj}$. The eigenvalues of L are integrals of motion and determine the asymptotic velocities.

The GCTC derived in [10–12, 14, 18] is also a completely integrable model. It allows Lax representation just like the standard real Toda chain [9, 29, 31] $\tilde{L}_t = [\tilde{A}.\tilde{L}]$, where:

$$\tilde{L} = \sum_{s=1}^{N} \left(\tilde{b}_s E_{ss} + \tilde{a}_s (E_{s,s+1} + E_{s+1,s}) \right), \quad A = \sum_{s=1}^{N} \left(\tilde{a}_s (E_{s,s+1} - E_{s+1,s}) \right), \tag{20}$$

where $\tilde{a}_s = m_{0s} e^{i\sigma_s} a_s$, $b_s = \mu_s + i\nu_s$. Like for the scalar case, the eigenvalues of $\tilde{L}$ are integrals of motion. If we denote by $\zeta_s = \kappa_s + i\eta_s$ (resp. $\tilde{\zeta}_s = \tilde{\kappa} + i\tilde{\eta}_s$) the set of eigenvalues of L (resp. $\tilde{L}$) then their real parts κ_s (resp. $\tilde{\kappa}_s$) determine the asymptotic velocities for the soliton train described by CTC (resp. GCTC). Thus, starting from the set of initial soliton parameters we can calculate $L|_{t=0}$ (resp. $\tilde{L}|_{t=0}$), evaluate the real parts of their eigenvalues and thus determine the asymptotic regime of the soliton train.

Regime (i). $\kappa_k \neq \kappa_j$ (resp. $\tilde{\kappa}_k \neq \tilde{\kappa}_j$) for $k \neq j$, i.e., the asymptotic velocities are all different. Then we have asymptotically separating, free solitons, see also [4, 16, 17, 24].

Regime (ii). $\kappa_1 = \kappa_2 = \cdots = \kappa_N = 0$ (resp. $\tilde{\kappa}_1 = \tilde{\kappa}_2 = \cdots = \tilde{\kappa}_N = 0$), i.e., all N solitons move with the same mean asymptotic velocity, and form a "bound state."

Regime (iii). A variety of intermediate situations when one group (or several groups) of particles move with the same mean asymptotic velocity; then they would form one (or several) bound state(s) and the rest of the particles will have free asymptotic motion.

Remark 1. *The sets of eigenvalues of L and $\tilde{L}$ are generically different. Thus varying only the polarization vectors one can change the asymptotic regime of the soliton train.*

Let us consider several particular cases.

Case 1. $\vec{n}_1 = \cdots = \vec{n}_N$. Since the vector $\vec{n}_1$ is normalized, then all coefficients $m_{ok} = 1$ and $\sigma_k = 0$. Then the interactions of the vector and scalar solitons are identical.

Case 2. $(\vec{n}^{\dagger}_{s+1}, \vec{n}_s) = 0$. Then the GCTC splits into two unrelated GCTC: one for the solitons $\{1, 2, \ldots, s\}$ and another for $\{s+1, s+2, \ldots . N\}$. If the two sets of soliton parameters are such that both groups of solitons are in bound state regimes, then we have two bound states.

Case 3. $\langle n^{\dagger}_{k+1} | \vec{n}_k \rangle = m_0 e^{i\varphi_0}$ — effective change of distance and phases of solitons. In this case we can rewrite $\tilde{a}_s = \exp((\tilde{Q}_{s+1} - \tilde{Q}_s)/2)$, where:

$$\tilde{Q}_{s+1} - \tilde{Q}_s = Q_{s+1} - Q_s + \ln m_0 + i\varphi_0, \tag{21}$$

i.e., the distance between any two neighboring vector solitons has changed by $\ln(m_0/2\nu_0)$; similar changes have the phases.

3. Asymptotic regimes for N-soliton trains with N $\geqslant 4$

The asymptotic regimes for scalar solitons and for small values of N are known for long time now, see [16, 17, 24]. Obviously for $N = 2$ we have only two possibilities: BSR and FAR. For $N = 3$ for the first time there appears MAR when two of the solitons form a bound state while the third one goes away off them. For $N > 3$ there were only fragmentary results, see the quasi-equidistant propagation of solitons in [16].

For the Manakov solitons formally the method is the same. The idea to use the integrability of CTC in order to develop a tool for the analysis of asymptotic behavior of N-soliton trains was developed in [10, 12, 14, 18]. Roughly speaking we have to use the characteristic polynomial of L_N whose generic form is:

$$P(z) = \det(L_N - z\mathbb{1}) = \sum_{k=0}^{N} p_k(\vec{a}, \vec{b}) z^k = \prod_{k=1}^{N} (z - z_k). \tag{22}$$

Next we have to analyze the roots z_k and formulate the conditions on the soliton parameters for which

$$\text{i) } \operatorname{Re}\ z_k = 0; \qquad \text{ii) } \operatorname{Im}\ z_k = 0. \tag{23}$$

Formally condition i) in (23) ensures the BSR, while condition ii) in (23) is responsible for the FAR.

However each soliton now has 6 parameters, so 3, 4 and 5 solitons will be parametrized by 18, 24 and 30 parameters respectively. The large number of parameters makes it difficult to derive explicit analytical results, or to do an exhaustive numerical studies. Of course some configurations of Manakov solitons behave just like the scalar ones. This happens if all $\vec{n}_k$ are equal. Naturally our aim is consider more interesting cases and demonstrate the important role that the polarization vectors play for the soliton interactions. Indeed m_{0k} in (17) take any value from 0 to 1, i.e., they 'regulate' the strength of the interaction. In particular, if the polarization vectors of two neighboring solitons are orthogonal, then they do not interact. In addition the phases σ_k modify the phase difference of the solitons which is a substantial factor in their interaction.

Situations when we have 2, 3 and 4 solitons are easier because we can write down explicit formulae for z_k in terms of the soliton parameters in the generic case. For two and three solitons most of this analysis for scalar solitons were done [16,17,24]. For bigger values of N such formulas are not done even for the scalar case, in which the number of the soliton parameters are $4N$. For $N = 4$ already the formulae for z_k are involved; in addition the number of the parameters is $4N = 16$. Therefore for $N \geqslant 4$ even for the scalar case only special configurations of soliton parameters are known. They are related to special choices of the soliton parameters that simplify the characteristic polynomial so that it reduces to, say a biquadratic equation. In addition, when it comes to Manakov solitons, the number of the parameters becomes $6N$.

Our aim here will be: first to revisit the particular cases considered before and, second, to propose special soliton configurations responsible for the BSR and FAR for any number of solitons. We will illustrate our results by several figures.

3.1. Asymptotic regimes for Manakov solitons

Let us now outline some effective ways of choosing soliton parameters that would ensure given asymptotic behavior of the solitons. The soliton parameters of the Manakov N-soliton train are

$6N$ and detailed study of the regions in which the solitons will develop given asymptotic regime does not seem possible. However we will outline several ways to effectively pick up configurations ensuring BSR or FAR asymptotic regimes.

Let us also remind several important issues that one needs to consider. First we need to specify what we will consider as asymptotic state. Obviously we need a criterium that would ensure us that we are in the asymptotic region. In our case we have two scales: $\epsilon^{1/2}$ and ϵ that are fundamental for the adiabatic approximation. It is reasonable to assume that the asymptotic times must be of the order of $1/\epsilon$. Our choices of soliton parameters are such that $\epsilon \simeq 10^{-2}$. So one could expect that the asymptotic times would be of the order of $\epsilon^{-1} \simeq 100$. At the same time we extend our numerics to about $t_{\rm as} \simeq 1000$ and in most cases we find good match between the CTC prediction and the numerics of Manakov model during all that period. This means that CTC models the Manakov model much better that we can expect. We can see from the figures presented here and from many others that we have done that the match could be much better.

Indeed, let us assume that we know how to split the 30-dimensional space of our soliton parameters into regions that correspond to the different asymptotic regimes. Obviously, if we choose the soliton parameters to be close to the 'border' between two different regimes we can expect that we would have a 'transition' area between the regimes, so the deviation from the CTC model will come up sooner than 1000. This is what we can see in Figs. 1, 2. In the right panel of Fig. 3 for $t \gg 300$ we see that the bound state of 5 solitons in fact transforms into a MAR. It 'peels off' the first and the fifth solitons that go freely away, and the other three still stay in a BSR. It seems that choosing the difference between the amplitudes stabilizes the BSR.

The general criterium that ensures FAR or BSR is based on the following well known proposition coming from linear algebra.

Proposition 1. *Let L_0 be symmetric $L_0 = L_0^T$ matrix with real-valued matrix elements. Then its eigenvalues z_{0j} will be real and different, i.e., $z_{0j} \neq z_{0k}$ for $k \neq j$.*

Corollary 1. *Let L_1 be symmetric (not hermitian) $L_1 = L_1^T$ matrix with purely imaginary matrix elements. Then its eigenvalues z_{1j} will be purely imaginary and different, i.e., $z_{1j} \neq z_{1k}$ for $k \neq j$.*

Proof. Follows directly from the Proposition if we consider $L_1 = iL_0$. □

In addition below we will assume that $\nu_0 = 0.5$ and $\mu_0 = 0$.

3.2. Generic FAR configurations

These configurations are characteristic for the real Toda chain solved by Moser [9, 29, 31].

In what follows we choose the polarization vectors $\vec{n}_k$ by setting:

$$\theta_k = \frac{k\pi}{13}, \qquad \gamma_k = \frac{k\pi}{g_0}. \tag{24}$$

where $g_0 = 8$, or $g_0 = 9$.

For the CTC using the Proposition we obtain:

$$\text{Im } b_k|_{t=0} = 0, \qquad \text{Im } a_k|_{t=0} = 0, \tag{25}$$

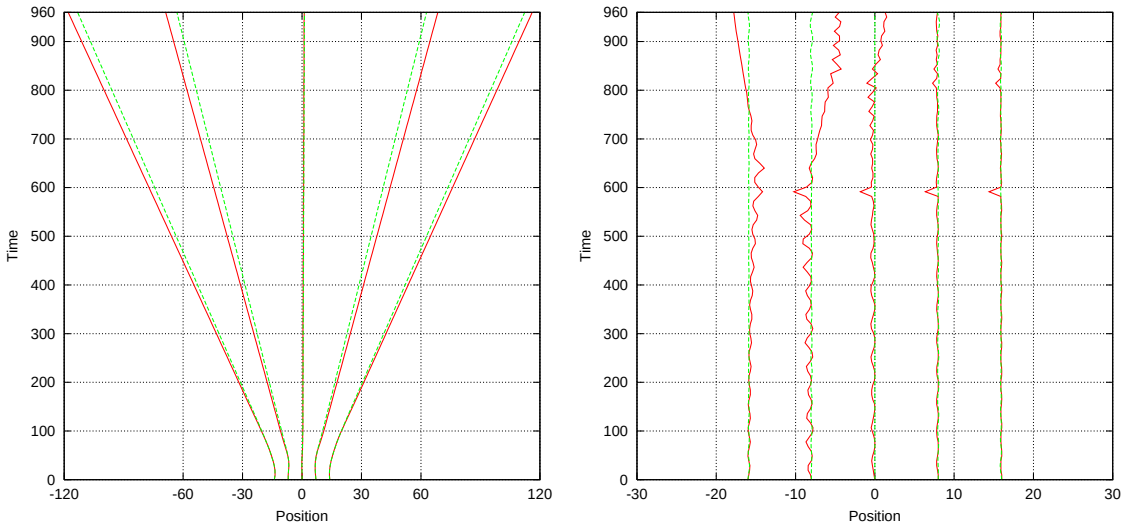


Fig. 1. Left panel: FAR with initial conditions $r_0 = 7.0$, $\mu_{00} = 0.01$, $\nu_{00} = 0.0$, $g_0 = 9$; Right panel: BSR for t up to 600 and MAR for $t > 700$ with initial conditions $r_0 = 8.0$, $\mu_{00} = 0.0$, $\nu_{00} = 0.05$, $g_0 = 9$. The rest of the parameters are defined by eqs. (26) and (28) respectively

which means that

$$\begin{aligned} &\nu_k|_{t=0} = 0.5, \qquad b_k|_{t=0} = \mu_k|_{t=0} = \mu_{0k}, \quad \theta_k = \frac{k\pi}{13}, \quad \gamma_k = \frac{k\pi}{g_0}, \\ &\xi_{0k} = (k-3)r_0, \quad \mu_{0k} = (k-3)\mu_{00}, \qquad \nu_{0k} = 0.5 + (k-3)\nu_{00}, \\ &\delta_{0,1} = 0, \qquad \delta_{0,k+1} - \delta_{0,k} = \sigma_k. \end{aligned} \tag{26}$$

Indeed, from the Proposition the eigenvalues of L will be real and different, which is FAR. A particular case of (26) as configuration ensuring FAR for scalar solitons was noticed long ago, namely choosing solitons with equal amplitudes (i.e., $\Delta\nu_k = 0$) and and out-of phase $\delta_{k+1} - \delta_k = \pi$ [4]. However, eq. (26) provides more general configurations, in which the solitons may have non-vanishing initial velocities, see Fig. 1.

3.3. Generic BSR configurations

Here we use the Corollary and impose on L the conditions:

$$\text{Re } b_k|_{t=0} = 0, \qquad \text{Re } a_k|_{t=0} = 0, \tag{27}$$

which means that

$$\begin{aligned} &b_k|_{t=0} = i\nu_k|_{t=0} = i\nu_{0k}, \qquad \theta_k = \frac{k\pi}{13}, \qquad \gamma_k = \frac{k\pi}{g_0}, \\ &\xi_{0k} = (k-3)r_0, \qquad \mu_{0k} = 0.0, \qquad \nu_{0k} = 0.5 + (k-3)\nu_{00}, \\ &\delta_{0,1} = 0, \qquad \delta_{0,k+1} - \delta_{0,k} = \sigma_k + \pi, \end{aligned} \tag{28}$$

This is also rather general and simple condition on the soliton parameters that fixes the initial velocities to be 0, but does not put restrictions (except the adiabatic ones) on the amplitudes and on the initial positions of the solitons.

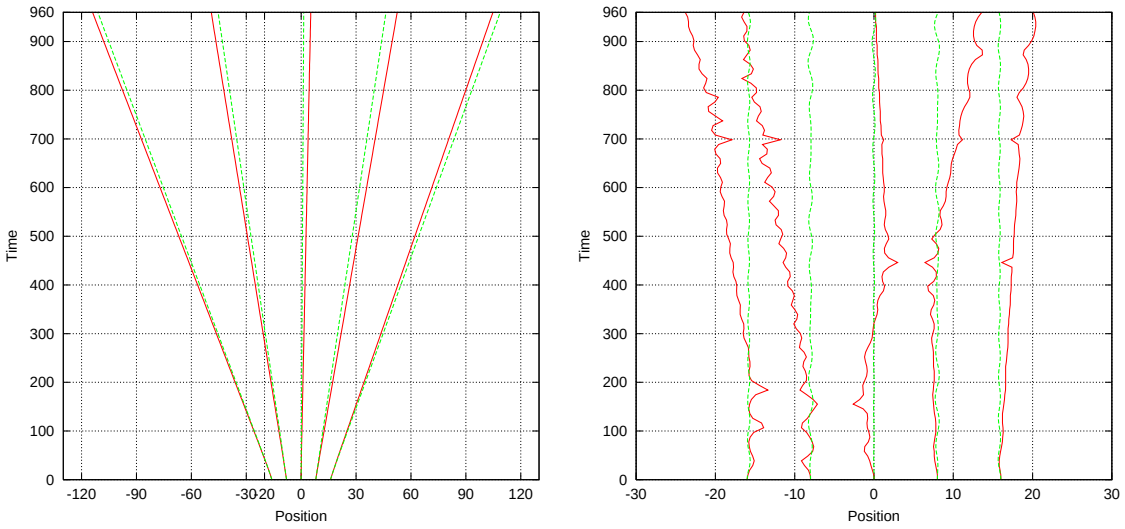


Fig. 2. Left panel: FAR with initial conditions $r_0 = 8.0$, $\mu_{00} = 0.02$, $\nu_{00} = 0.0$, $g_0 = 4$; Right panel: BSR for t up to 300 and MAR for $t > 500$ with initial conditions $r_0 = 8.0$, $\mu_{00} = 0.0$, $\nu_{00} = 0.03$, $g_0 = 4$. The rest of the parameters are defined by eqs. (26) and (28) respectively

3.4. Symmetric configurations of soliton parameters

In addition to these we find other configurations of soliton parameters that provide FAR or BSR. To this end we use special symmetric constraints on L described below. These constraints will leave only one of ν_{0k} and a_{0k} independent. As a result the characteristic polynomial of L will factorize and we will find that all roots are proportional to each other.

Let us give few examples of them. We will provide the corresponding Lax matrix, its characteristic polynomial and eigenvalues.

- $N = 3$, $P_3 = z(z^2 - 4(a^2 + b^2))$:

$$L_3 = \begin{pmatrix} b & \sqrt{2}a & 0 \\ \sqrt{2}a & 0 & \sqrt{2}a \\ 0 & \sqrt{2}a & -b \end{pmatrix}, \qquad (29)$$
$$z_{1,2} = \pm 2\sqrt{a^2 + b^2}, \qquad z_3 = 0;$$

- $N = 4$, $P_4 = (z^2 - a^2 - b^2)(z^2 - 9(a^2 + b^2))$

$$L_4 = \begin{pmatrix} 3b & \sqrt{3}a & 0 & 0 \\ \sqrt{3}a & b & 2a & 0 \\ 0 & 2a & -b & \sqrt{3}a \\ 0 & 0 & \sqrt{3}a & -3b \end{pmatrix}, \qquad (30)$$
$$z_{1,2} = \pm\sqrt{a^2 + b^2}, \qquad z_{3,4} = \pm 3\sqrt{a^2 + b^2};$$

- $N = 5$, $P_5 = z(z^2 - a^2 - b^2)(z^2 - 4(a^2 + b^2))$

$$L_5 = \begin{pmatrix} 2b & \sqrt{3}a & 0 & 0 & 0 \\ \sqrt{2}a & b & 2a & 0 & 0 \\ 0 & 2a & 0 & \sqrt{3}a & 0 \\ 0 & 0 & \sqrt{3}a & -b & \sqrt{2}a \\ 0 & 0 & 0 & \sqrt{2}a & -2b \end{pmatrix} \tag{31}$$

$$z_{1,2} = \pm\sqrt{a^2 + b^2}, \qquad z_{3,4} = \pm 2\sqrt{a^2 + b^2}, \qquad z_5 = 0;$$

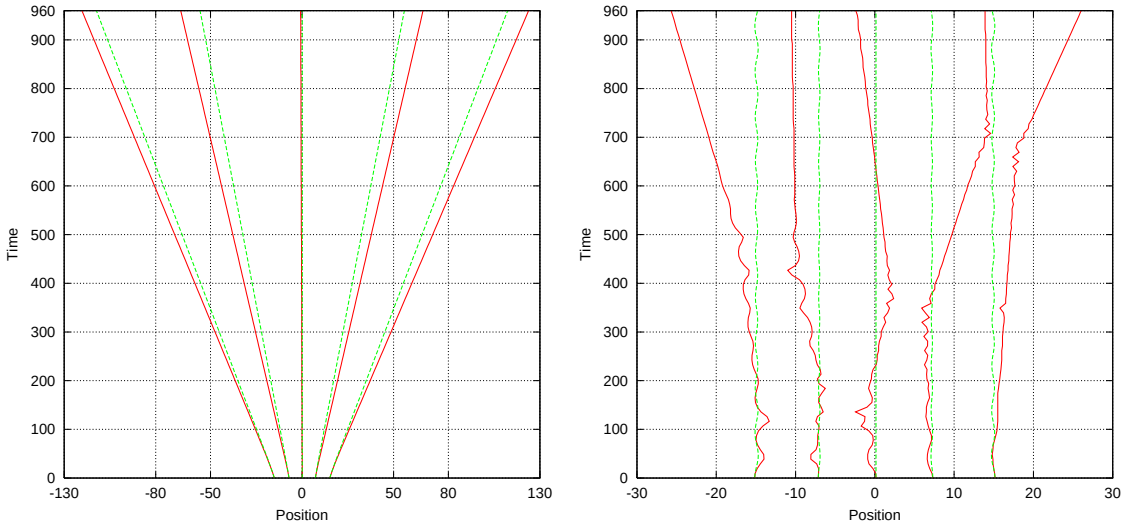


Fig. 3. Left panel: FAR with initial conditions $\mu_{00} = 0.02$, $\nu_{00} = 0.0$, $g_0 = 4$; Right panel: BSR for t up to 300 and MAR for $t > 400$ with initial conditions $\mu_{00} = 0.0$, $\nu_{00} = 0.03$, $g_0 = 4$. The rest of the parameters are defined by eqs. (37) and (38) respectively

- $N = 6$, $P_6 = (z^2 - a^2 - b^2)(z^2 - 9(a^2 + b^2))(z^2 - 25(a^2 + b^2))$:

$$L_6 = \begin{pmatrix} 5b & \sqrt{5}a & 0 & 0 & 0 & 0 \\ \sqrt{5}a & 3b & \sqrt{3}a & 0 & 0 & 0 \\ 0 & \sqrt{8}a & b & 3a & 0 & 0 \\ 0 & 0 & 3a & -b & \sqrt{8}a & 0 \\ 0 & 0 & 0 & \sqrt{8}a & -3b & \sqrt{5}a \\ 0 & 0 & 0 & 0 & \sqrt{5}a & -5b \end{pmatrix}, \tag{32}$$

$$z_{1,2} = \pm\sqrt{a^2 + b^2}, \qquad z_{3,4} = \pm 3\sqrt{a^2 + b^2}, \qquad z_{5,6} = \pm 5\sqrt{a^2 + b^2}.$$

Such examples can be found for any value of N; from algebraic point of view they are related to the the maximal embedding of $sl(2)$ as a subalgebra of $sl(N)$.

In order to ensure FAR or BSR we need to impose on a and b the condition that

$$\text{FAR} \qquad a^2 + b^2 > 0, \qquad \text{BSR} \qquad a^2 + b^2 < 0. \tag{33}$$

Initial conditions for BSR of 5 scalar solitons:

$$\begin{aligned}
&\xi_1 = -2r_0 + \frac{\ln 6}{2\nu_0}, \qquad \xi_2 = -r_0 + \frac{\ln 3}{2\nu_0}, \quad \xi_3 = 0, \qquad \xi_4 = r_0 - \frac{\ln 3}{2\nu_0}, \quad \xi_5 = 2r_0 - \frac{\ln 6}{2\nu_0},\\
&\nu_k = 0.5 + (3-k)\nu_{00}, \quad \mu_k = 0, \qquad \delta_k = k\pi, \qquad k = 1, \dots, 5.
\end{aligned} \tag{34}$$

Initial conditions for FAR of 5 scalar solitons:

$$\begin{aligned}
&\xi_1 = -2r_0 + \frac{\ln 6}{2\nu_0}, \quad \xi_2 = -r_0 + \frac{\ln 3}{2\nu_0}, \quad \xi_3 = 0, \qquad \xi_4 = r_0 - \frac{\ln 3}{2\nu_0}, \quad \xi_5 = 2r_0 - \frac{\ln 6}{2\nu_0},\\
&\nu_k = 0.5, \qquad \mu_k = (3-k)\mu_{00}, \quad \delta_k = \frac{k\pi}{2}, \qquad k = 1, \dots, 5.
\end{aligned} \tag{35}$$

For Manakov solitons the initial positions are determined by:

$$\begin{aligned}
&\xi_{10} = -2r_0 - \frac{1}{2\nu_0}\ln\frac{m_{01}m_{02}m_{03}m_{04}}{6}, \quad \xi_{20} = -r_0 - \frac{1}{2\nu_0}\ln\frac{m_{02}m_{03}m_{04}}{3m_{01}},\\
&\xi_{30} = -\frac{1}{2\nu_0}\ln\frac{m_{03}m_{04}}{m_{01}m_{02}},\\
&\xi_{40} = r_0 + \frac{1}{2\nu_0}\ln\frac{m_{01}m_{02}m_{03}}{3m_{04}}, \qquad \xi_{50} = 2r_0 + \frac{1}{2\nu_0}\ln\frac{m_{01}m_{02}m_{03}m_{04}}{6}.
\end{aligned} \tag{36}$$

For the numerics we again fix the polarization vectors as in (24) and evaluate ξ_{0k} by the formula (36). The result are given in Tab. 1 and 2 below.

In order to have FAR we choose the amplitudes, velocities and the phases of the solitons by:

$$\begin{aligned}
&\nu_k = 0.5, \qquad \mu_k = (k-3)\mu_{00}, \qquad k = 1, 2, \dots, 5,\\
&\delta_{10} = 0, \qquad \delta_{20} = \delta_{10} + \sigma_1 + \pi, \qquad \delta_{30} = \delta_{10} + \sigma_1 + \sigma_2 + \pi,\\
&\delta_{40} = \delta_{30} + \sigma_1 + \sigma_2 + \sigma_3 + \pi, \qquad \delta_{50} = \delta_{40} + \sigma_1 + \sigma_2 + \sigma_3 + \sigma_4 + \pi.
\end{aligned} \tag{37}$$

For the BSR we choose the amplitudes, velocities and the phases of the solitons by:

$$\begin{aligned}
&\nu_k = 0.5 + (k-3)\nu_{00}, \qquad \mu_k = 0, \qquad k = 1, 2, \dots, 5,\\
&\delta_{10} = 0, \qquad \delta_{20} = \delta_{10} + \sigma_1, \qquad \delta_{30} = \delta_{10} + \sigma_1 + \sigma_2,\\
&\delta_{40} = \delta_{30} + \sigma_1 + \sigma_2 + \sigma_3, \qquad \delta_{50} = \delta_{40} + \sigma_1 + \sigma_2 + \sigma_3 + \sigma_4.
\end{aligned} \tag{38}$$

3.5. Numeric values for the intial parameters

In Tabs. 1 and 2 we list the numeric values for m_{0k} and σ_k for the two typical choices of θ_k and γ_k used above.

Table 1. Initial phases for Fig. 1 and Fig. 2

δ_{0k}	left panel	right panel
$k=1$	0.0	0.0
$k=2$	2.868037	–0.273554
$k=3$	–0.405708	–0.405708
$k=4$	2.781038	–0.360554
$k=5$	–0.150741	–0.150741

δ_{0k}	left panel	right panel
$k=1$	0.0	0.0
$k=2$	2.484841	–0.656751
$k=3$	–1.006917	–1.006917
$k=4$	2.258187	–0.883405
$k=5$	–0.354039	–0.354039

Table 2. Initial phases and positions for Fig. 3

	left panel		right panel	
	δ_{0k}	ξ_{0k}	δ_{0k}	ξ_{0k}
$k=1$	0.0	–15.154654	0.0	–15.154654
$k=2$	2.484841	–7.133487	–0.656751	–7.133487
$k=3$	–1.006917	0.140982	–1.006917	0.140982
$k=4$	2.258187	7.305540	–0.883405	7.305540
$k=5$	–0.354039	15.154654	–0.354039	15.154654

4. Conclusions and discussion

The above analysis can be extended to any number of solitons. As we mentioned above, the symmetric Lax matrices are realizations of the maximal embedding of the $sl(2)$ algebra as a subalgebra of $sl(N)$. In this case we effectively reduce the N-soliton interactions to an effective 2-soliton interactions. Therefore the symmetric configurations studied above allow only two asymptotic regimes: BSR and FAR. We make the hypothesis that it would be possible to construct more general symmetric Lax matrices that would be responsible for effective 3-soliton interactions. In this paper we included numerical tests only for 5 soliton interactions. However previously we have run test starting with 2-solitons and ending with 9-soliton configurations. Our results are that the CTC models adequately not only the purely solitonic interactions, but also the effects of external potentials and other perturbations on them.

An interesting question is how long should we wait for the asymptotic regime. This question is directly related to the other one: What are the limits of applicability of CTC? In our simulations we have chosen $\varepsilon_0 \simeq 0.01$ which means that the asymptotic time must be of the order of $1/\varepsilon_0 \simeq 100$. At the same time in a number of cases we find good match between the CTC and the numeric solutions of Manakov model even until 1 000. This is what we see in our tests in this paper for the free asymptotic regimes (left panels of all figures). The situation is different for the bound state regimes. While in Fig. 1 we see good match until about 700, in Figs. 1 and 3 the good match goes until 300. After that the trajectories of CTC keep to the BSR, but some of the real solitons 'escape away' after that. However in all cases we find that CTC provides good descriptions until times about three times larger than the asymptotic one.

MDT was supported by Fulbright – Bulgarian-American Commission for Educational Exchange under Grant No 19-21-07.

References

[1] F.Kh.Abdullaev, B.B.Baizakov, S.A.Darmanyan, V.V.Konotop, M.Salerno, Nonlinear excitations in arrays of Bose-Einstein condensates, *Phys. Rev. A*, **64**(2001), 043606.

[2] G.P.Agrawal, Nonlinear Fiber Optics, Academic, San Diego, 1995 (2nd edn).

[3] D.Anderson, M.Lisak, Bandwidth limits due to mutual pulse interaction in optical soliton communication systems, *Optics Lett.*, **11**(1986), no. 3, 174–176. DOI: 10.1364/OL.11.000174

[4] J.M.Arnold, Complex Toda lattice and its application to the theory of interacting optical solitons, *JOSA A*, **15A**(1998), no. 5, 1450–1458. DOI: 10.1364/JOSAA.15.001450

[5] J.M.Arnold, Stability of solitary wave trains in Hamiltonian wave systems, *Phys. Rev. E*, **60**(1999), no. 1, 979–986. DOI: 10.1103/PhysRevE.60.979

[6] R.Carretero-Gonzales, V.S.Gerdjikov, M.D.Todorov, N-soliton interactions: Effects of linear and nonlinear gain/loss, *AIP CP*, **1895**(2017), 040001.

[7] R.Carretero-Gonzalez, K.Promislow, Localized breathing oscillations of Bose-Einstein condensates in periodic traps, *Phys. Rev. A*, **66**(2002), 033610.

[8] E.V.Doktorov, V.S.Shchesnovich, Modified nonlinear Schrödinger equation: Spectral transform and N-soliton solution, *J. Math. Phys.*, **36**(1995), 7009. DOI: 10.1063/1.531204

[9] H.Flaschka, The Toda lattice. II. Existence of integrals, *Phys. Rev. B*, 1924, **9**(1974).

[10] V.S.Gerdjikov, N-Soliton interactions, the Complex Toda chain and stability of NLS soliton trains, In: Prof. E. Kriezis (Ed), Proceedings of the International Symposium on Electromagnetic Theory, vol. 1, 307–309 (Aristotle University of Thessaloniki), Greece, 1998.

[11] V.S.Gerdjikov, Complex Toda chain – an integrable universal model for adiabatic N-soliton interactions, In: M. Ablowitz, M. Boiti, F. Pempinelli, B. Prinari (Eds), "Nonlinear Physics: Theory and Experiment. II", World Scientific, 2003.

[12] V.S.Gerdjikov, Modeling soliton interactions of the perturbed vector nonlinear Schrödinger equation, *Bulgarian J. Phys.*, **38**(2011), 274–283.

[13] V.S.Gerdjikov, B.B.Baizakov, M.Salerno, Modelling adiabatic N-soliton interactions and perturbations, *Theor. Math. Phys.*, **144**(2005), no. 2, 1138–1146.

[14] V.S.Gerdjikov, E.V.Doktorov, N.P.Matsuka, N-soliton train and generalized Complex Toda chain for Manakov system, *Theor. Math. Phys.*, **151**(2007), no. 3, 762–773.

[15] V.S.Gerdjikov, E.V.Doktorov, J.Yang, Adiabatic interaction of N ultrashort solitons: Universality of the Complex Toda chain model, *Phys. Rev. E*, **64**(2001), 056617.
DOI: 10.1103/PhysRevE.64.056617

[16] V.S.Gerdjikov, E.G. Evstatiev, D.J.Kaup, G.L.Diankov, I.M.Uzunov, Stability and quasi-equidistant propagation of NLS soliton trains, *Phys. Lett. A*, **241**(1998), 323–328.

[17] V.S.Gerdjikov, D.J.Kaup, I.M.Uzunov, E.G.Evstatiev, Asymptotic behavior of N-soliton trains of the Nonlinear Schrödinger equation, *Phys. Rev. Lett.*, **77**(1996), 3943–3946.
DOI: 10.1103/PhysRevLett.77.3943

[18] V.S.Gerdjikov, N.A.Kostov, E.V.Doktorov, N.P.Matsuka, Generalized Perturbed Complex Toda chain for Manakov system and exact solutions of the Bose-Einstein mixtures, *Mathematics and Computers in Simulation*, **80**(2009), 112–119.
DOI: 10.1016/j.matcom.2009.06.013

[19] V.S.Gerdjikov, A.V.Kyuldjiev, M.D.Todorov, Manakov solitons and effects of external potential wells, DCDS Supplement, Vol. 2015, 2015, 505–514.

[20] V.S.Gerdjikov, M.D.Todorov, On the effects of sech-like potentials on Manakov solitons, AIP Conference Proceedings, Vol. 1561, 2013, 75–83. DOI: 10.1063/1.4827216

[21] V.S.Gerdjikov, M.D.Todorov, A.V.Kyuldjiev, Polarization effects in modeling soliton interactions of the Manakov model, AIP Conference Proceedings, Vol. 1684, 080006 (2015). DOI: 10.1063/1.4934317

[22] V.S.Gerdjikov, M.D.Todorov, A.V.Kyuldjiev, Adiabatic interactions of Manakov soliton – effects of cross-modulation. Special Issue of Wave Motion — "Mathematical modeling and physical dynamics of solitary waves: From continuum mechanics to field theory", I.C. Christov, M.D. Todorov, S. Yoshida (Eds), Wave Motion, **71**(2017), 71–81.

[23] V.S.Gerdjikov, I.M.Uzunov, Adiabatic and non-adiabatic soliton interactions in nonlinear optics, *Physica D*, **152-153**(2001), 355–362. DOI: 10.1016/S0167-2789(01)00179-8

[24] V.S.Gerdjikov, I.M.Uzunov, E.G.Evstatiev, G.L.Diankov, Nonlinear Schrödinger equation and N-soliton interactions: Generalized Karpman-Solov'ev approach and the Complex Toda chain, *Phys. Rev. E*, **55**(1997), no. 5, 6039–6060.

[25] V.I.Karpman, Soliton evolution in the presence of perturbation, *Physica Scripta*, **20**(1979), 462–478.

[26] V.I.Karpman, V.V.Solov'ev, A perturbational approach to the two-soliton systems, *Physica D*, **3**(1981), no. 3, 487–502.

[27] Yu.S.Kivshar, B.A.Malomed, Dynamics of solitons in nearly integrable systems, *Rev. Mod. Phys.*, **61**(1989), no. 4, 763–915.

[28] S.V.Manakov, On the theory of Two-dimensional stationary self-focusing electromagnetic waves, *Sov. Phys. JETP*, **38**(1974), 248–253.

[29] S.V.Manakov, On the complete integrability and stochastization in discrete dynamical systems, *Sov. Phys. JETP*, **40**(1974) 269–274.

[30] M.Midrio, S.Wabnitz, P.Franco, Perturbation theory for coupled nonlinear Schrödinger equations, *Phys. Rev. E*, **54**(1996), 5743.

[31] J.Moser, In: Dynamical Systems, Theory and Applications. Lecture Notes in Physics, Vol. 38, Springer Verlag, 1975.

[32] V.S.Shchesnovich, E.V.Doktorov, Perturbation theory for the modified nonlinear Schrцdinger solitons, *Physica D*, **129**(1999), 115.

[33] M.D.Todorov, V.S.Gerdjikov, A.V.Kyuldjiev, Multi-soliton interactions for the Manakov system under composite external potentials, *Proc. Estonian Academy of Sciences. Phys.-Math. Series*, **64**(2015), no. 3, 368—378.

[34] I.M.Uzunov, M.Gölles, F.Lederer, Stabilization of soliton trains in optical fibers in the presence of third-order dispersion, *JOSA B*, **12**(1995), no. 6, 1164.

[35] V.E.Zakharov, S.V.Manakov, S.P.Novikov, L.P.Pitaevskii, Theory of Solitons: The Inverse Scattering Method, Plenum, N.Y., Consultants Bureau, 1984.

[36] V.E.Zakharov, A.B.Shabat, Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, *Zh. Eksp. Teor. Fiz.*, **61**(1971), 118–134.

Об асимптотическом поведении N-солитонных последовательностей Манакова в адиабатическом приближении

Владимир С. Герджиков
Национальный исследовательский ядерный университет "МИФИ"
Москва, Российская Федерация
Институт математики и информатики Болгарской академии наук
София, Болгария
Институт перспективных физических исследований, Новый болгарский университет
София, Болгария
Михаил Д. Тодоров
Государственный университет Сан-Диего
Сан-Диего, Калифорния, США
Технический университет Софии
София, Болгария

Аннотация. Мы анализируем динамическое поведение N-солитонных последовательностей Манакова в адиабатическом приближении. Эволюция этих солитонных последовательностей моделируется комплексной цепочкой Тода (КЦТ), которая является вполне интегрируемой динамической системой. Вычисляя собственные значения ее матрицы Лакса мы можем определить асимптотическую скорость каждого из солитонов. Это позволяет нам описать конфигурации солитонных параметров при которых солитонная последовательность переходит в каждом из двух основных ясимптотических режимов: (а) режим связанного состояния и (б) режим асимптотически свободного поведения. В частности мы нашли явное описание специальных симметрических конфигураций N солитонов которые обеспечивают как, режим связанного состояния, так и режим асимптотически свободного поведения. Мы установили отличное совпадение между траекториями, предсказываемых КЦТ с теми, которые получаются при численном решении модели Манакова для широкого класса солитонных параметров. Это подтверждает справедливость нашей модели.

Ключевые слова: модель Манакова, солитонные взаимодействия, адиабатическое приближение, комплексная цепочка Тода.

DOI: 10.17516/1997-1397-2020-13-6-694-707
УДК 517.958:519.633

On the Construction of Solutions to a Problem with a Free Boundary for the Non-linear Heat Equation

Alexander L. Kazakov*
Matrosov Institute for System Dynamics and Control Theory SB RAS
Irkutsk, Russian Federation
Lev F. Spevak†
Institute of Engineering Science Ural Branch RAS
Ekaterinburg, Russian Federation
Ming-Gong Lee‡
Chung Hua University
Hsinchu City, Taiwan

Received 08.06.2020, received in revised form 14.07.2020, accepted 10.08.2020

Abstract. The construction of solutions to the problem with a free boundary for the non-linear heat equation which have the heat wave type is considered in the paper. The feature of such solutions is that the degeneration occurs on the front of the heat wave which separates the domain of positive values of the unknown function and the cold (zero) background. A numerical algorithm based on the boundary element method is proposed. Since it is difficult to prove the convergence of the algorithm due to the non-linearity of the problem and the presence of degeneracy the comparison with exact solutions is used to verify numerical results. The construction of exact solutions is reduced to integrating the Cauchy problem for ODE. A qualitative analysis of the exact solutions is carried out. Several computational experiments were performed to verify the proposed method.

Keywords: non-linear heat equation, heat wave, boundary element method, approximate solution, exact solution, existence theorem.

Citation: A.L. Kazakov, L.F. Spevak, M.-G. Lee, On the Construction of Solutions to a Problem with a Free Boundary for the Nonlinear Heat Equation, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 694–707. DOI: 10.17516/1997-1397-2020-13-6-694-707.

Introduction

We consider the non-linear parabolic heat equation [1] with a source (sink)

$$T_t = \Delta\Psi(T) + \Phi(T), \tag{1}$$

which is also called the generalized porous medium equation [2]. If $\Psi(0) = 0$ and $\Phi(T)$ is power function Eq. (1) can be written as

$$u_t = u\Delta u + \gamma(\nabla u)^2 + \alpha u^\beta. \tag{2}$$

*kazakov@icc.ru https://orcid.org/0000-0002-3047-1650
†lfs@imach.uran.ru https://orcid.org/0000-0003-2957-6962
‡mglee1990@gmail.com https://orcid.org/0000-0001-9405-2247

Here, $\gamma > 0$, $\alpha, \beta \neq 0$ are constants, $\alpha > 0$ means the presence of a source and $\alpha < 0$ corresponds to a sink. In what follows Cartesian coordinates are used.

Various forms of equation (2) are used to describe processes in continuum mechanics [2, 3], plasma physics [1], etc. This mathematical object has a distinctive property related to the propagation of perturbations with finite velocity, which is not typical for parabolic equations.

For equation (2) the problem of initiating a heat wave is considered. The heat wave is a construction consisting of two hypersurfaces: $u(t, \mathbf{x}) \geqslant 0$ and $u(t, \mathbf{x}) \equiv 0$ that continuously joined along some sufficiently smooth manifold $\Gamma(t, \mathbf{x}) = 0$. The latter determines the front of the heat wave. Since the front is unknown in advance and it is determined simultaneously with the construction of the unknown function, we have a special problem with a free boundary [3], where $u|_{\Gamma(t,\mathbf{x})=0} = 0$. The boundary conditions have the form

$$u|_{b(\mathbf{x})=0} = f(t, \mathbf{x}), \quad f(0, \mathbf{x}) = 0, \tag{3}$$

where $b(\mathbf{x}), f(t, \mathbf{x})$ are sufficiently smooth functions.

Previously, problem (2), (3) was already considered in the case $\alpha = 0$, i.e., without a source (sink). Solutions were constructed both in the form of special series [4] and with the use of the boundary element method (BEM) [5]. In this paper, we propose an approximate method for constructing solutions to the problem of heat wave initiation based on the BEM.

There are various approaches to solve boundary value problems for parabolic equations using the BEM. The most natural one is to use the BEM based on time-dependent fundamental solutions [6]. This method is not suitable for solving problem (2), (3) because of the non-linearity of the right-hand side and because of the presence of a movable boundary (the heat wave front). Therefore, it is preferable to use a time-stepping BEM [7], where the boundary value problem for the elliptic equation is considered at each step, and the fundamental solution of the Laplace equation is used. In addition to the classical BEM, it is used in the method of fundamental solutions (MFS) [8], as well as in the interior field method (IFM) [9] and the null field method (NFM) [10] in annular domains for the case of circular symmetry. For problem (2), (3), we obtain the Poisson equation with a non-linear right-hand side at each time step. The dual reciprocity boundary element method (DRBEM) [11] is most suitable for solving this equation.

We did not establish the convergence of the developed method. Therefore, to verify numerical results exact solutions in the form of travelling wave are used [2]. Construction of exact solutions is reduced to the solution of the Cauchy problem for a second-order ordinary differential equation with a singularity.

Finding exact solutions of non-linear partial differential equations is an important field of modern mathematics. There is a wide variety of methods to find exact solutions. Among these methods we emphasize the group analysis method that was proposed and developed by L. V. Ovsiannikov and his colleagues [12, 13]. A review of methods for constructing exact solutions to equations of mathematical physics can be found, for example, in handbook [14]. Various generalizations and modifications of the method of separation of variables [15, 16] are especially often used to construct exact solutions of non-linear parabolic equations having the form (1).

1. Formulation of the problem

In the case of one spatial variable, equation (2) can be written as

$$u_t = u u_{\rho\rho} + \gamma u_\rho^2 + \frac{\nu u_\rho}{\rho} + \alpha u^\beta. \tag{4}$$

Here, $\nu = 0, 1, 2$ corresponds to Cartesian, cylindrical and spherical coordinate systems, respectively; $\rho = \sqrt{\sum_{i=1}^{\nu+1} x_i^2}$, where x_i are the Cartesian coordinates. Condition (3) has the form

$$u|_{\rho=R} = f(t), \quad f(0) = 0, \tag{5}$$

where $R \geqslant 0$ is some constant which must obviously be positive for $\nu \neq 0$.

The cases $\nu = 0$ and $\nu = 1$ are considered here. It follows from previous results [17] that for Cartesian and cylindrical coordinate systems problem (4), (5) has unique analytical heat-wave type solution (in the form of a convergent Taylor series). The coefficients of the series are determined from the solution of systems of linear algebraic equations. However, the radius of convergence of the series is usually small, and it can be estimated only in some spacial cases. To solve this problem and obtain an approximate solution to the problem of heat wave initiating at a given time interval $[0, t^*]$, we usually use a step-by-step method based on the boundary element approach [5]. We also note that presence of additional term (source) requires significant modification of the previously developed approach.

2. Solution algorithm based on BEM

Problem (4), (5) is solved in the specified time interval $t \in [0, t^*]$ by a step-by-step method based on the BEM. At each time step $t_k = kh$, where h is the step size, we solve the spatial problem obtained from (4), (5) with $t = t_k$. The solution domain where the unknown function u is positive is the interval $\rho \in [0, a(t_k))$, and $\rho = a(t)$ is the equation of motion of the heat wave front, $u(t_k, a(t_k)) = 0$.

For certainty, it is assumed that the heat wave front moves from the origin. Since $a(t_k)$ is unknown in advance, the solution domain is also unknown at the moment $t = t_k$. That is why, we interchange the desired function and the spatial variable ρ [5]. Equation (4) takes the following form

$$\rho_t \rho_u^2 = u\rho_{uu} - \gamma\rho_u - \frac{\nu u \rho_u^2}{\rho} + \alpha u^\beta \rho_u^3. \tag{6}$$

We rewrite equation (6) in the form of the Poisson equation and obtain at $t = t_k$ the boundary value problem

$$\rho_{uu} = F(u, \rho, \rho_t, \rho_u), \quad \rho|_{u=L} = R, \tag{7}$$

where $F(u, \rho, \rho_t, \rho_u) = (\rho_t \rho_u^2 + \gamma\rho_u)/u + \nu\rho_u^2/\rho + \alpha u^{\beta-1}\rho_u^3$, $\rho = \rho(t_k, u)$ is the unknown function and $L = f(t_k)$. The unknown heat wave front for the original problem is defined by the condition $\rho|_{u=0} = a(t_k)$.

At the front of the heat wave we have [18]

$$q^{(\rho)}|_{u=0} = \left.\frac{\partial\rho}{\partial n}\right|_{u=0} = \frac{\gamma}{a'(t_k)}, \tag{8}$$

where $q^{(\rho)}$ is the flow of $\rho(t_k, u)$, n is the external normal to the boundary of the solution domain, $n(0) = -1$, $n(L) = 1$. It follows from the results presented in [17] that $a'(t_k) \neq 0$.

Thus, we arrive to the boundary value problem (7), (8) in the domain $u \in [0, L]$. Using the boundary element method, we write the solution of this problem in the following form

$$\rho(v) = q_{1k}^{(\rho)} u^*(v, 0) + q_{2k}^{(\rho)} u^*(v, L) - \rho_{1k} q^*(v, 0) - R q^*(v, L) - \int_0^L F(u, \rho, \rho_t, \rho_u)\, u^*(v, u) du. \tag{9}$$

Here, $v \in (0, L)$, $\rho_{1k} = \rho(t_k, 0)$, $\rho_{2k} = \rho(t_k, L) = R$, $q_{1k}^{(\rho)} = q^{(\rho)}(t_k, 0)$, $q_{2k}^{(\rho)} = q^{(\rho)}(t_k, L)$, $u^*(v, u)$ is the fundamental solution of the one-dimensional stationary problem, $q^*(v, u) = \partial u^*(v, u)/\partial n$ [6]. Values of ρ_{1k}, $q_{1k}^{(\rho)}$ and $q_{2k}^{(\rho)}$ are not specified by the boundary conditions and should be determined.

Taking the limits $v \to 0$ and $v \to L$ in equation (9), we obtain the system of two boundary integral equations

$$\rho_{1k} - q_{1k}^{(\rho)} L = Q_1, \quad \rho_{1k} + q_{2k}^{(\rho)} L = Q_2, \tag{10}$$

where $Q_1 = R - \int\limits_0^L F\left(u, \rho, \rho_t, \rho_u\right) u^*(0, u) du$, $Q_2 = R + \int\limits_0^L F\left(u, \rho, \rho_t, \rho_u\right) u^*(L, u) du$.

Using quadratic approximation of the function $\rho(t, 0) = a(t)$ on the interval $[t_{k-1}, t_k]$, one can write condition (8) in the following approximate form

$$\rho_{1k} - \rho_{1(k-1)} = \frac{\gamma h \left(q_{1k}^{(\rho)} + q_{1(k-1)}^{(\rho)}\right)}{2 q_{1k}^{(\rho)} q_{1(k-1)}^{(\rho)}}. \tag{11}$$

Expressing ρ_{1k} from (11) and substituting it into the first equation (10), we obtain an equation in the unknown $q_{1k}^{(\rho)}$

$$L \left(q_{1k}^{(\rho)}\right)^2 + \left(Q_1 - \rho_{1(k-1)} - \frac{\gamma h}{2 q_{1(k-1)}^{(\rho)}}\right) q_{1k}^{(\rho)} + \frac{\gamma h}{2} = 0. \tag{12}$$

Since the derivative of the unknown function in (4), (5) is negative along the front $u_\rho|_{\rho=a(t)} < 0$ (for chosen direction of the movement of the heat wave front), the inverse function obeys the inequality $\rho_u|_{u=0} = 1/(u_\rho|_{u=0}) < 0$. Therefore, $q_{1k}^{(\rho)} = n(0)\rho_u|_{u=0} = -\rho_u|_{u=0} > 0$, and a suitable solution of equation (12) has the form

$$q_{1k}^{(\rho)} = \frac{1}{2L}\left[\rho_{1(k-1)} - Q_1 + \frac{\gamma h}{2 q_{1(k-1)}^{(\rho)}} + \sqrt{\left(Q_1 - \rho_{1(k-1)} - \frac{\gamma h}{2 q_{1(k-1)}^{(\rho)}}\right)^2 + 2\gamma h L}\right]. \tag{13}$$

Substituting (13) into (10), we can find ρ_{1k} and $q_{2k}^{(\rho)}$:

$$\rho_{1k} = q_{1k}^{(\rho)} L + Q_1, \quad q_{2k}^{(\rho)} = \frac{Q_2 - \rho_{1k}}{L}. \tag{14}$$

Since function $F(u, \rho, \rho_t, \rho_u)$ in (7) depends on the unknown function and its derivatives, we use the following iterative procedure to solve problem (7), (8). Let us take $\rho^{(0)} \equiv R$ and $Q_1 = 0$, $Q_2 = 0$ as the initial values. Then the i-th iteration of the solution has the form

$$\begin{gathered} \rho^{(i)}(v) = q_{1k}^{(\rho)(i)} u^*(v, 0) + q_{2k}^{(\rho)(i)} u^*(v, L) - \\ -\rho_{1k}^{(i)} q^*(v, 0) - R q^*(v, L) - \int_0^L F(u, \rho^{(i-1)}, \rho_t^{(i-1)}, \rho_u^{(i-1)}) u^*(v, u) du. \end{gathered} \tag{15}$$

The boundary values of the unknown function and flow can be found according to (13), (14) as

$$q_{1k}^{(\rho)(i)} = \frac{1}{2L}\left[\rho_{1(k-1)} - Q_1^{(i-1)} + \frac{\gamma h}{2 q_{1(k-1)}^{(\rho)}} + \sqrt{\left(Q_1^{(i-1)} - \rho_{1(k-1)} - \frac{\gamma h}{2 q_{1(k-1)}^{(\rho)}}\right)^2 + 2\gamma h L}\right],$$

$$\rho_{1k}^{(i)} = q_{1k}^{(\rho)(i)} L + Q_1^{(i-1)}, \quad q_{2k}^{(\rho)(i)} = (Q_2^{(i-1)} - \rho_{1k}^{(i)})/L.$$

Here, $Q_1^{(i-1)}$, $Q_1^{(i-1)}$ are calculated from the previous iteration:

$$Q_1^{(i-1)} = R - \int_0^L F^{(i-1)} u^*(0,u)du, \quad Q_2^{(i-1)} = R + \int_0^L F^{(i-1)} u^*(L,u)du, \tag{16}$$

where $F^{(i-1)} = F\left(u, \rho^{(i-1)}, \rho_t^{(i-1)}, \rho_u^{(i-1)}\right)$. The iteration process is terminated at the n-th iteration if $\left|(\rho_{1k}^{(n)} - \rho_{1k}^{(n-1)})/\rho_{1k}^{(n)}\right| < \varepsilon$, where ε is a given constant. Then the approximate solution of (7), (8) at $t = t_k$ is $\rho(t_k, u) = \rho^{(n)}(u)$. Since this solution is continuous, the solution of problem (4), (5) at $t = t_k$, $u(t_k, \rho)$ can be determined without a loss of accuracy.

The developed algorithm allows us to construct a solution that is continuous with respect to a spatial variable for problem with a free boundary (4), (5) at each given time step.

To calculate integrals $\int\limits_0^L F\left(u, \rho, \rho_t, \rho_u\right) u^*(v,u)du$ in (15), (16), we use the dual reciprocity method [11] based on the expansion of $F\left(u, \rho, \rho_t, \rho_u\right)$ in terms of radial basis functions (RBF)

$$F\left(u, \rho, \rho_t, \rho_u\right) = \sum_{k=1}^{n} \alpha_k \phi_k(u). \tag{17}$$

Functions ϕ_k depend on the distance between the current point and collocation points $u_1, u_2, \ldots, u_n$ that belong to the interval $[0, L]$: $\phi_k(x) = \phi\left(r_k\right)$, where $r_k = |u - u_k|$. For each function ϕ_k there is a function $\widehat{w}_k$ such that $\phi_k = \Delta\widehat{w}_k$. After substituting expansion (17) into the integrand and twice integrating by parts, we obtain the following equality

$$\begin{gathered} \int_0^L F\left(u, \rho, \rho_t, \rho_u\right) u^*(v,u)du = \\ = \sum_{k=1}^{n} \alpha_k[-\widehat{w}_k(v) + \widehat{p}_k(0)u^*(v,0) + \widehat{p}_k(L)u^*(v,L) - \widehat{w}_k(0)q^*(v,0) - \widehat{w}_k(L)q^*(w,L)], \end{gathered} \tag{18}$$

where $\widehat{p}_k(x) = \partial\widehat{w}_k(u)/\partial n$. The coefficients α_k for each iteration are determined from the system of equations obtained from (17) for the current iteration $\rho^{(i)}(u)$ at the collocation points

$$F\left(u_j, \rho^{(i)}\left(u_j\right), \rho_t^{(i)}\left(u_j\right), \rho_u^{(i)}\left(u_j\right)\right) = \sum_{k=1}^{n} \alpha_k \phi_k(u_j), j = 1, 2, \ldots, n.$$

The use of the simplest functions $\phi_k = r_k$ as RBFs [5, 18] results in stable convergence of iterative processes and good accuracy of solutions. However, it is rather complicated to use these functions to solve problems with the source term because convergence of iterative processes is unstable and depends on the parameters of the problem. Obviously, the additional non-linear term requires a more precise expansion.

It is difficult to analyse the influence of RBFs on convergence analytically. Therefore, we perform a numerical analysis of the influence of used RBFs on convergence and accuracy of the solution. We consider linear function $\phi_k = 1 + r_k$, polyharmonic spline $\phi_k = r_k^3$ and multi-quadric function $\phi_k = \sqrt{1 + \epsilon r_k^2}$. Stable convergence is observed when the last two functions are used. At the same time, multi-quadric function ensures better accuracy of the solution so we use it in calculations. The results are shown in Section 4.

3. Construction and study of exact solutions

It is problematic to prove convergence of the algorithm based on the BEM. One possible way is to construct and study exact solutions and then use them to verify numerical results.

Plain geometry. We consider the non-linear heat equation with a source for $\nu = 0$

$$u_t = uu_{xx} + \gamma u_x^2 + \alpha u^\beta. \tag{19}$$

We assume that in this case the heat wave front is defined by following equation

$$u(t,x)|_{x=a(t)} = 0. \tag{20}$$

If functions $a(t)$ and u^β are analytical ($\beta \in \mathbb{N}$) then it follows from the previously proved theorems (see, for example, [16]) that problem (19), (20) has unique analytical solution in form of a power series with respect to variable $z = x - a(t)$ with recurrently determined coefficients. We consider the case when the condition of the analyticity of the source is not satisfied. Let $\beta > 0$, $\beta \in \mathbb{R}$ in equation (19). For non-integer β the source function can not be expanded into a Taylor series with respect to powers of u. We construct the solution in the form of a travelling wave $u = v(z)$, $z = x - \mu t - \eta$, $\mu > 0$, $\eta \geqslant 0$. Due to the invariance of equation (19) one can take $\eta = 0$.

Substituting $v(z)$ in (19), we obtain the following ordinary differential equation

$$vv'' + \gamma(v')^2 + \mu v' + \alpha v^\beta = 0. \tag{21}$$

Solving this equation with the condition $v(0) = 0$, we obtain the solution of problem (19), (20) which is a heat wave with the front $x = a(t) = \mu t + \eta$ (if it exists). Properties of solutions of this type for $\beta \in \mathbb{N}$ were previously studied [16]. For non-integer values of β, as far as we know, this problem was not previously considered and it is now studied for the first time.

In this case one cannot impose arbitrary condition for the derivative at $z = 0$. If we set $z = 0$ in both parts of Eq. (21) then we obtain the quadratic equation

$$\gamma(v'(0))^2 + \mu v'(0) = 0,$$

which has two roots $v'(0) = -\mu/\gamma$ and $v'(0) = 0$. For other values of $v'(0)$ equation (21) is incompatible. Thus, we obtain from the continuity condition that either

$$v(0) = 0, \quad v'(0) = -\mu/\gamma, \text{ or} \tag{22}$$

$$v(0) = 0, \quad v'(0) = 0. \tag{23}$$

Problems (21), (22) and (21), (23) have specific properties that they inherit from original problem (19), (20). In particular, at the starting point $z = 0$ function $v(z)$ (that is the term at the higher derivative) turns to zero. It means that problems cannot be written in normal form, and the classical theorems on existence of solutions of the Cauchy problem are not applicable. It is obvious that (21), (23) has trivial solution $v \equiv 0$. Special study is required to prove the existence and uniqueness of the solution of (21), (22) and the existence of a non-trivial solution of (21), (23).

In what follows we consider only the case $\alpha > 0$. It means that there is an influx of energy (matter) into the system. This case is widely encountered in applications.

Theorem 1. *The Cauchy problem* (21), (22) *for* $\alpha > 0$, $\beta \geqslant 1$ *has unique solution* $v(z) \in C_{[z_0,0]} \cap C^2_{(z_0,0)}$ *for some* $z_0 < 0$.

Proof. Using substitution $v' = p$, we lower the order of equation (21). Then it takes the form

$$vp\frac{dp}{dv} + \gamma p^2 + \mu p + \alpha v^\beta = 0. \tag{24}$$

Using the change of variable $w = v^\beta$ in equation (24), we obtain

$$\beta w\frac{dp}{dw} + \gamma p + \mu + \frac{\alpha w}{p} = 0. \tag{25}$$

Since for $\beta > 0$ the equality $v^\beta|_{v=0} = 0$ holds then for equation (24) conditions (22) correspond to

$$p(0) = -\mu/\gamma. \tag{26}$$

Let us perform a qualitative analysis of equation (21). To do this, we consider the equivalent dynamic system

$$\frac{dw}{d\xi} = \frac{\beta}{\mu}wp, \quad \frac{dp}{d\xi} = -\frac{\gamma}{\mu}p^2 - p - \frac{\alpha}{\mu}w. \tag{27}$$

Here the parametrization is performed in such a way that $dz = \mu w d\xi$. System (27) is very close to system (4.3) from [19]. The first equations differ by a positive constant multiplier on the right-hand side. The second equations are the same.

System (27) has two singular points $M_1(0, -\mu/\gamma)$ and $M_2(0, 0)$. It follows from qualitative analysis [19] that M_1 has the topological type "saddle", and M_2 is the "saddle-node" with one nodal and two saddle sectors. Considering condition (26), we are primarily interested in phase trajectories that enter point M_1 and/or leave it. In this case, there is no need to consider the left phase half-plane $w < 0$ (except $\beta = k/m$, where k, m are natural odd numbers) because $w = v^\beta$.

There is a separatrix S that tends to the point M_1 as $\xi \to +0$. Phase trajectories which are located to the right of S bypass the nodal sector bounded by S and the coordinate axis Op. The phase trajectories inside the nodal sector tend to M_2 as $\xi \to +\infty$ (Fig. 1).

The separatrix S corresponds to the solution $v = v_*(z)$ of problem (21), (22). Taking into account the conditions of the theorem, the solution has the following properties: 1) v is defined and continuous on some interval $[z_0, 0]$; 2) v is twice continuously differentiable on the interval $(z_0, 0)$; 3)@$v_*(0) = v_*(z_0) = 0$, $v_*(z) \geqslant 0$; 4) $v'_*(z)$ changes sign once, $v'_*(z_{\max}) = 0$, $v_*(z_{\max}) = v_{\max}$, $\lim\limits_{z \to z_0+0} v'(z) = +\infty$; 5) $v''_*(z) < 0$. Let us note that condition $\beta \geqslant 1$ guarantees that properties 1 and 2 hold. The schematic representation of function $v(z)$ is shown in Fig.2.

It seems impossible to find the exact values $z_*, z_{\max}, v_{\max}$. Further we discuss how to find interval estimates for them.

Remark 1. *For* $\beta = k/m$, *where* k, m *are natural odd numbers,* $k > m$, *the left phase half-plane* $w = v^\beta < 0$ *can also be considered. The solution of problem* (21), (22) *can be continued to the right from the point* $z = 0$ *as well as problem* (21), (23) *at* $z < 0$ *has a non-trivial solution. Both solutions are negative so they are meaningless from the physical point of view.*

Interval estimates are constructed for the key parameters of the solution of (21), (23) in the particular case $\beta = 1$ to simplify the study. We follow the procedure suggested in [19].

Let us use linear substitution of the unknown function and independent variable

$$\tilde{z} = Az, \ \tilde{v} = Bv, \quad A = \alpha\gamma/[\mu(\gamma + 1)], \ B = \alpha\gamma^2/[\mu^2(\gamma + 1)]. \tag{28}$$

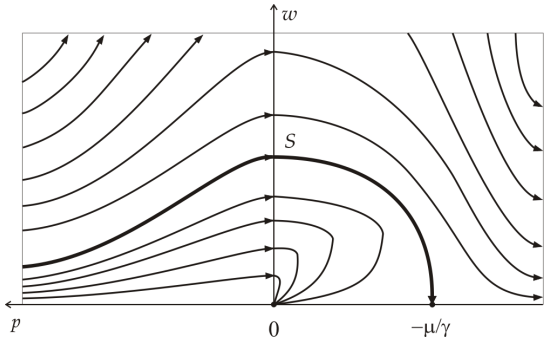


Fig. 1. Phase portrait

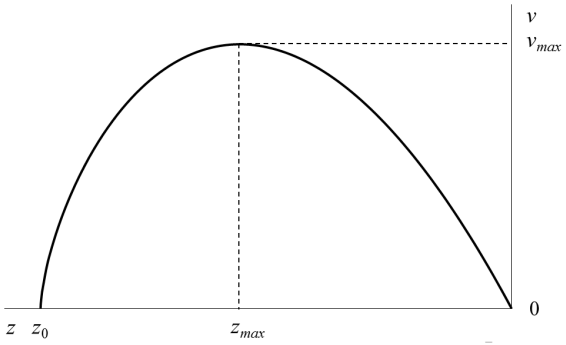


Fig. 2. The traveling wave configuration

Then problem (21), (23) takes the form

$$vv'' + \gamma(v')^2 + \gamma v' + (\gamma+1)v = 0, \quad v(0) = 0,\ v'(0) = -1. \tag{29}$$

Here the tilde is omitted.

We analyse the properties of solution of (29) and find estimates for $z_*, z_{\max}, v_{\max}$. In the proof of Theorem 1 we show that for $z \in [z_{\max}, 0]$ function v decreases, $-1 \leqslant v' \leqslant 0$ and $v'(z_{\max}) = 0$. Then from (29) we have that $-1-\gamma \leqslant v'' \leqslant -1$, and $v''(0) = -1$, $v''(z_{\max}) = -\gamma - 1$.

Integrating the upper bound for v'' on the interval $[z, 0]$ and taking into account the Cauchy conditions (29), we obtain that

$$v' \leqslant -(\gamma+1)z - 1, \quad v \geqslant -0.5(\gamma+1)z^2 - z,\ z_{\max} \leqslant z \leqslant 0. \tag{30}$$

It follows from the first inequality (30) that $z_{\max} \leqslant -1/(\gamma+1)$. The right-hand side of the second inequality at $z_1 = -1/(\gamma+1) \geqslant z_{\max}$ takes the maximum value that satisfies the inequality $v(z_1) \geqslant 1/[2(\gamma+1)]$. Since $v(z_1) \leqslant v_{\max}$ we obtain that $v_{\max} \geqslant 1/[2(\gamma+1)]$.

Integrating the lower bound for v'' on the interval $[z, 0]$ we obtain that

$$v' \geqslant -z - 1, \quad v \leqslant -0.5z^2 - z,\ 0 < z \leqslant z_{\max}. \tag{31}$$

It follows from (31) that $z_{\max} \geqslant -1$ and $v_{\max} \leqslant 0.5$.

When $z \in [z_0, z_{\max}]$ function v increases, i.e., $v' \geqslant 0$. Taking into account this inequality, it follows from (29) that $v'' \leqslant -\gamma - 1$. Integrating it twice, we obtain $v_{\max} - v \geqslant (\gamma+1)(z_{\max}-z)^2/2$, $z_0 \leqslant z \leqslant z_{\max}$. Since $v_{\max} \leqslant 0.5$, for $z = z_0$ we have $0 < z_{\max} - z_0 \leqslant 1/\sqrt{\gamma+1}$ whence it follows that $-1/\sqrt{\gamma+1} - 1 \leqslant z_0 \leqslant -1/(\gamma+1)$. So, we have all the required estimates. Let us apply the transformation inverse to (28). Then we obtain the following inequalities

$$-\frac{\mu(\gamma+1)}{\alpha\gamma} \leqslant z_{\max} \leqslant -\frac{\mu}{\alpha\gamma};\ -\frac{\mu}{\alpha\gamma}(\sqrt{\gamma+1}+\gamma+1) \leqslant z_0 < z_{\max};\ \frac{\mu^2}{2\alpha\gamma^2} \leqslant v_{\max} \leqslant \frac{\mu^2(1+\gamma)}{2\alpha\gamma^2}. \tag{32}$$

We omit cumbersome transformations that are required to construct estimates of the form (32) for the general case. Now a solution of equation (19) can be found. The solution has the form of a heat wave that propagates with the constant velocity $u(t,x) = v_*(x - \mu t - \eta)$.

An interesting feature of this solution is that it is a soliton (solitary wave). We point out that if $\alpha = 0$ (no source) then solution has explicit form $u = -\mu(x - \mu t - \eta)/\gamma$. It is easy to show that it is not a soliton.

Derivation of travelling wave solution. Since one can not obtain analytical solution of problem (21), (22) we solve it numerically. It follows from the proof of Theorem 1 that solution should be constructed on the interval $[z_0, 0]$, and $z_0 < 0$ is unknown. We only know that $v(z_0) = 0$ and $v'(z) \to \infty$ when $z \to z_0$. This is the problem with a free boundary, and its formulation is non-standard for solving by the boundary element method. Note that we prefer to use the BEM because, unlike difference methods, it allows one to construct a continuous solution.

For convenience, we use substitution $V(z) = v(-z)$ and resolve equation (21) with respect to the higher derivative. Taking into account (22), we have the Cauchy problem

$$V'' = \frac{1}{V}\left[\mu V' - \gamma\left(V'\right)^2 - \alpha V^{\beta}\right],\ V(0) = 0,\ q(0) = -\frac{\mu}{\gamma}. \tag{33}$$

It is difficult to find its solution with satisfactory accuracy on the interval $z \in [0, z_*]$, where $z_* = -z_0$, because of specific properties (mentioned above) of the unknown function. A correct solution can be found in two stages. In the first stage, problem (33) is solved on the interval $z \in [0, L]$, where $L < z_*$ is selected in such a way that $V'(L) < 0$, i.e., $L > -z_{\max}$. Then the BEM iterative procedure is used to obtain continuous solution of this problem (see [20]).

One cannot construct a solution of problem (33) on the interval $z \in [L, z_*]$ in the original formulation because the derivative is unbounded. Then, in the second stage, we interchange the independent variable and the unknown function, just as we did in Section 2. As a result, we obtain the inverse Cauchy problem for the unknown function $z(V)$.

$$z'' = \frac{z'}{V}\left[\gamma - \mu z' + \alpha V^{\beta}(z')^2\right],\ z(L^*) = L,\ q_z(L^*) = \frac{1}{q(L)}, \tag{34}$$

where $V \in [0, L^*]$, $q_z = \partial z/\partial n$, $L^* = V(L)$; $V(L)$ and $q(L)$ are obtained in the first stage.

Now problem (34) can be solved with the use of the iterative BEM. As a result, in particular, the value $z(0) = z_*$ is found. The continuousness of the found function allows us to determine $V(z)$ for $z \in [0, z_*]$ and the solution of problem (21), (22) without loss of accuracy. Note that estimates (32) can be used to select parameter L.

Cylindrical geometry. For $\nu = 1$, there are no exact travelling wave solutions for equation (4). It is known [16] that for $\beta = 1$ equation (4) has a self-similar solution $u(t, \rho) = a(t)a'(t)v(r)$, where $r = \rho/(Re^{\theta t})$, and $R > 0, \theta$ are constant. Function $v(r)$ satisfies the Cauchy problem

$$vv'' + \gamma(v')^2 + \left(r + \frac{v}{r}\right)v' + \left(\frac{\alpha}{R} - 2\right)v = 0, \quad v(1) = 0,\ v'(1) = -\frac{1}{\gamma}. \tag{35}$$

This solution corresponds to a heat wave with an exponential law of motion of the heat wave front $a(t) = Re^{\theta t}$. Since equation (35) is not autonomous, it is difficult to perform a qualitative analysis in this case, and we don't consider this task here.

In conclusion, we note that solution of problem (35) by the iterative boundary element method [20] does not require additional modification because the unknown function in the solution domain $r \in (0, 1]$ monotonically decreases. In this case, to solve problem (4), (5) in the time interval $t \in [0, t_*]$, it is sufficient to solve problem (35) in the segment $r \in [a(0)/a\left(t_*\right), 1]$.

4. Computational experiment

Verification of the BEM algorithm. Here we test the BEM algorithm developed in Section 2 by comparing the calculation results with the exact solutions presented in Section 3 for various values of parameters.

Example 1. We consider the problem in the case of plane geometry ($\nu = 0$) with $\gamma = 1/3, \alpha = 1, \mu = 1, \eta = 0$. Parameter β is varied: $\beta = 0.8;\ 1;\ 1.2;\ 1.4;\ 1.6;\ 1.8;\ 2$.

The accuracy of solving problem (4), (5) using the BEM algorithm is tested as follows. The exact travelling wave solution $u^e(t, x)$ is found by solving problem (21), (22) using the boundary element method described in Section 3. When the exact solution is found condition (5) that corresponds to this solution is

$$u|_{x=0} = u^e(t, 0). \tag{36}$$

Next, numerical solution of problem (4), (36) is calculated using BEM. Then it is compared with the exact solution $u^e(t, x)$. Numerical and exact solutions for $\beta = 1.4$ are shown in Fig. 3.

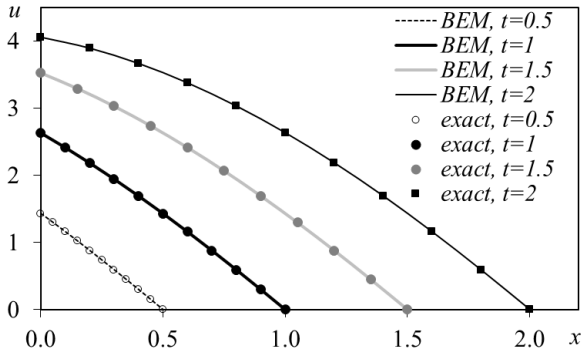


Fig. 3. Travelling heat wave for $\beta = 1.4$

To estimate the accuracy of the numerical solution we compare the law of motion of the heat wave front found by the BEM and the law of motion $x = \mu t + \eta$ of the exact solution. The relative error in determining the heat wave front is shown in Tab. 1.

Table 1. The relative error in determining the heat wave front

t	$\beta = 0.8$	$\beta = 1$	$\beta = 1.2$	$\beta = 1.4$	$\beta = 1.6$	$\beta = 1.8$	$\beta = 2$
0.5	4.96E–03	1.77E–03	6.29E–04	2.10E–04	7.22E–05	2.19E–05	2.86E–05
1	4.63E–03	1.61E–03	5.85E–04	2.10E–04	8.88E–05	4.97E–05	4.37E–05

The results show that the accuracy of the solution decreases as parameter β decreases. The results are not acceptable for $\beta < 0.8$. This seems to be related to the fact that the term $u^{\beta-1}$ (see (7)) has a singularity at $u = 0$ if $\beta < 1$. Note that Theorem 1 is also valid only for $\beta \geqslant 1$. Nevertheless, one should note that the acceptable numerical results are obtained not only under the conditions of Theorem 1 but also for $0 \ll \beta < 1$.

Example 2. We consider the problem in the case of cylindrical geometry ($\nu = 1$) with $\gamma = 1/3$, $\alpha = 1$, $\beta = 1$, $R = 1$, $\theta = 1$.

The exact solution $u^e(t, \rho)$ is found by solving problem (35) using the BEM [20]. The boundary condition for problem (4), (5) is

$$u|_{\rho=R} = u^e(t, R). \tag{37}$$

Comparison of the BEM solution of problem (4), (37) and the exact solution is shown in Fig. 4.

Thus, the results demonstrate the effectiveness of the developed algorithm for solving the problem of heat wave initiating for a non-linear heat equation with a source.

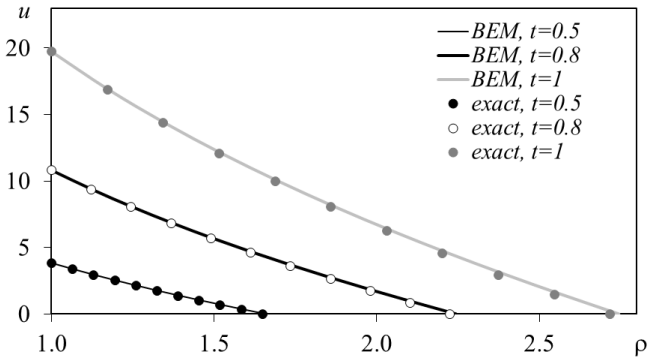


Fig. 4. Heat wave with exponential front

Traveling wave. The second part of the computational experiment is devoted to the numerical analysis of estimates obtained for the parameters of the solution of problem (21), (22). Tab. 2 shows values of $z_*, z_{\max}$ and $v_{\max}$, as well as boundaries of their interval estimates $z_- \leqslant z_{\max} \leqslant z_+$, $z^- \leqslant z_0 \leqslant z_{\max}$, $v_- \leqslant v_{\max} \leqslant v_+$ (see (32)) for $\alpha = \mu = 1$ and various γ. One can see that the estimates are relatively accurate for $z_{\max}$ and $v_{\max}$. The values of $z_{\max}$ are closer to the left border of the interval, and the values of $v_{\max}$ are closer to the right border. For the wave length $z_* = -z_0$ the rough estimate was obtained, and further improvement is needed.

The results of calculations show that obtained analytical estimates can be probably improved. So, for all calculations performed (both presented in Tab. 2 and not included in it) the inequalities $|z_{\max} - z_M|/|z_M| < 0.05$, $v_- + 0.7\ \Delta v < v_{\max} < v_- + 0.8\ \Delta v$ are valid. Here, $z_M = (z_- + z_+)/2$, $\Delta v = v_+ - v_-$. It follows from (28) and (29) that if we have values of $z_0(\gamma, \alpha, \mu)|_{\alpha=\mu=1}$, $z_{\max}(\gamma, \alpha, \mu)|_{\alpha=\mu=1}$ and $v_{\max}(\gamma, \alpha, \mu)|_{\alpha=\mu=1}$, we can find $z_0(\gamma, \alpha, \mu)$, $z_{\max}(\gamma, \alpha, \mu)$ and $v_{\max}(\gamma, \alpha, \mu)$ for all α and μ as $z_0(\gamma, \alpha, \mu) = = \mu\alpha^{-1} z_0(\gamma, 1, 1)$, $z_{\max}(\gamma, \alpha, \mu) = \mu\alpha^{-1} z_{\max}(\gamma, 1, 1)$, $v_{\max}(\gamma, \alpha, \mu) = \mu^2\alpha^{-1} v_{\max}(\gamma, 1, 1)$.

Table 2. Estimates of travelling wave parameters

γ	α	μ	z_-	$z_{\max}$	z_+	z^-	z_0	v_-	$v_{\max}$	v_+
1	1	1	–2	–1.547918	–1	–3.414214	–2.328672	0.5	0.858849	1
0.5	1	1	–3	–2.575608	–2	–5.449490	–4.265408	2	2.744421	3
1/3	1	1	–4	–3.590301	–3	–7.464102	–6.229110	4.5	5.643519	6
0.2	1	1	–6	–5.611311	–5	–11.477226	–10.194475	12.5	14.4864942	15

Conclusions

In this study we consider the problem of a heat wave initiating for a non-linear heat equation with a source and construct the solution on a specified finite time interval. We develop a step-by-step algorithm based on the iterative BEM using the dual reciprocity method. We choose

systems of radial basis functions that ensure convergence of the iterative process at each step of the solution. To verify the developed algorithm we use exact travelling wave solutions. Their construction is reduced to solving the Cauchy problem for the ODE with a singularity. For this Cauchy problem we prove the existence and uniqueness theorem of the classical solution that does not have to be analytical. A qualitative study of the solution properties was performed, and some estimates for the amplitude and wave length were obtained. To solve the Cauchy problem numerically, we develop an iterative algorithm based on the BEM. It allows one to determine correctly the boundary of the solution domain where the derivative of the unknown function tends to infinity. The performed calculations show the effectiveness and accuracy of the developed computational algorithm.

Further research can be directed towards expanding the proposed approach to other types of problems and to multidimensional equations. It is also interesting to apply new methods such as the method of fundamental solutions, the interior field method and the null field method to solve the considered elliptic equations and to compare these methods with the BEM.

The study was funded by RFBR (research project no. 20-07-00407) and by RFBR and MOST (research project no. 20-51-S52003.

References

[1] A.A.Samarskii, V.A.Galaktionov, S.P.Kurdyumov, A.P.Mikhailov, Blow-up in quasilinear parabolic equations, Berlin–New York, Walter de Gruyte, 1995.

[2] J.L.Vazquez, The Porous Medium Equation: Mathematical Theory, Clarendon Press, Oxford, 2007.

[3] V.K.Andreev, Yu.A.Gaponenko, O.N.Goncharova, V.V.Pukhnachev, Mathematical models of convection, Berlin, De Gruyter Studies in Mathematical Physics, 2012.

[4] M.Yu.Filimonov, L.G.Korzunin, A.F.Sidorov, Approximate methods for solving nonlinear initial boundary-value problems based on special construction of series, *Rus. J. Numer. Anal. Math. Modelling*, **8**(1993), no. 2, 101–125.

[5] A.L.Kazakov, O.A.Nefedova, L.F.Spevak, Solution of the problem of initiating the heat wave for a nonlinear heat conduction equation using the boundary element method, *Comp. Math. Math. Phys.*, **59**(2019), no. 6, 1015–1029. DOI: 10.1134/S0965542519060083

[6] P.K.Banerjee, R.Butterfield, Boundary Element Methods in Engineering Science, London, McGraw-Hill, 1981.

[7] M.Tanaka, T.Matsumoto, S.Takakuwa, Dual reciprocity BEM for time-stepping approach to the transient heat conduction problem in nonlinear materials, *Comput. Methods Appl. Mech. Eng.*, **195**(2006), 4953–4961.

[8] Z.C.Li, M.G.Lee, H.T.Huang, J.Y.Chiang, Neumann problems of 2D Laplace's equation by method of fundamental solutions, *Appl. Num. Math.*, **119**(2017), 126–145. DOI: 10.1016/j.apnum.2017.04.004

[9] Z.C.Li, J.Y.Chiang, H.T.Huang, M.G.Lee, The interior field method for Laplace's equation in circular domains with circular holes, *Eng. Anal. Boundary Elem.*, **67**(2016), 173–185. DOI: 10.1016/j.enganabound.2016.03.006

[10] M.G.Lee, Z.C.Li, L.Zhang, H.T.Huang, J.Y.Chiang, Algorithm singularity of the null-field method for Dirichlet problems of Laplace's equation in annular and circular domains *Eng. Anal. Boundary Elem.*, **41**(2014), 160–172. DOI: 10.1016/j.enganabound.2014.01.013

[11] L.C.Wrobel, C.A.Brebbia, D.Nardini, The dual reciprocity boundary element formulation for transient heat conduction, In: Finite elements in water resources VI, Berlin, Springer-Verlag, 1986, 801–811.

[12] L.V.Ovsiannikov, Group analysis of differential equations, Academic Press, New York–London, 1982.

[13] V.K.Andreev, O.V.Kaptsov, V.V.Pukhnachov, V.V.Rodionov, Applications of group-theoretical methods in hydrodynamics, Dordrecht, Kluwer Academic Publishers, 1998.

[14] A.D.Polyanin, V.F.Zaitsev, Handbook of nonlinear partial differential equations, Boca Raton-London-New York, Chapman and Hall/CRC, 2011.

[15] V.V.Pukhnachev, Exact multidimensional solutions of the nonlinear diffusion equation, *J. Appl. Mech. Technical Phys.*, **36**(1995), no. 2, 169–176.

[16] A.L.Kazakov, On exact solutions to a heat wave propagation boundary-value problem for a nonlinear heat equation, *Sib. Electronic Math. Reports*, **16**(2019), 1057–1068 (in Russian). DOI: 10.33048/semi.2019.16.073

[17] A.L.Kazakov, P.A.Kuznetsov, A.A.Lempert, On a Heat Wave for the Nonlinear Heat Equation: An Existence Theorem and Exact Solution, Continuum Mechanics, Applied Mathematics and Scientific Computing: Godunov's Legacy, 2020, Springer, 223–228. DOI: 10.1007/978-3-030-38870-6_29

[18] A.L.Kazakov, L.F.Spevak, Numerical and analytical studies of a nonlinear parabolic equation with boundary conditions of a special form, *Appl. Math. Model.*, **37**(2013), 6918–6928. DOI: 10.1016/j.apm.2013.02.026

[19] A.L.Kazakov, Sv.S.Orlov, S.S.Orlov, Construction and study of exact solutions to a nonlinear heat equation, *Sib. Math. J.*, **59**(2018), no. 3, 427–441. DOI: 10.1134/S0037446618030060 Study of Exact Solutions to A Nonlinear Heat Equation. Sib Math J 59, 427–441 (2018). https://doi.org/10.1134/S0037446618030060

[20] A.L.Kazakov, P.A.Kuznetsov, L.F.Spevak, Analytical and numerical construction of heat wave type solutions to the nonlinear heat equation with a source, *Journal of Mathematical Sciences*, **239**(2019), no. 1, 111–122. DOI: 10.1007/s10958-019-04294-x

О построении решений задачи со свободной границей для нелинейного уравнения теплопроводности

Александр Л. Казаков
Институт динамики систем и теории управления имени В.М. Матросова СО РАН
Иркутск, Российская Федерация
Лев Ф. Спевак
Институт машиноведения УрО РАН
Екатеринбург, Российская Федерация
Минг-Гонг Ли
Университет Чунг Хуа
Город Синьчжу, Тайвань

Аннотация. В статье обсуждается построение решений задачи со свободной границей для нелинейного уравнения теплопроводности, которые имеют тип тепловой волны. Особенностью таких решений является то, что уравнение имеет вырождение на фронте тепловой волны, который разделяет область положительных значений искомой функции и холодный (нулевой) фон. Предложен численный алгоритм решения указанной проблемы на основе метода граничных элементов. Поскольку доказать сходимость алгоритма не удается из-за нелинейности задач и наличия вырождения, в качестве метода верификации расчетов выбрано сравнение с точными решениями, построение которых сводится к интегрированию задачи Коши для ОДУ. Проведено качественное исследование последних. Выполнены иллюстрирующие расчеты, на основании которых с использованием результатов качественного анализа сделаны содержательные выводы.

Ключевые слова: нелинейное уравнение теплопроводности, тепловая волна, метод граничных элементов, приближенное решение, точное решение, теорема существования.

DOI: 10.17516/1997-1397-2020-13-6-708-717
УДК 536.46

A Coupled Mathematical Model for the Synthesis of Composites

Anna G. Knyazeva*
Natalia V. Bukrina†
Institute of Strength Physics and Materials Science SB RAS
Tomsk, Russian Federation

Received 10.07.2020, received in revised form 25.08.2020, accepted 20.09.2020

Abstract. The work proposes a model for synthesizing a composite "metallic matrix–reinforcing inclusions". The solution is based on two algorithms demonstrating similar results. It is shown that, like in classic models of combustion, there is a domain of model parameters where a transition to the stationary regime is possible. It is demonstrated that taking into account the thermal and mechanical processes alters the effective properties (thermal capacity and thermal effects of the reaction) and provokes the formation of a new heat source conditioned by the interaction of different physical processes.

Keywords: composite synthesis, pulsed heating, consecutive-parallel reactions, comparison of algorithms.

Citation: A.G. Knyazeva, N.V. Bukrina, A Coupled Mathematical Model for the Synthesis of Composites, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 708–717. DOI: 10.17516/1997-1397-2020-13-6-708-717.

Introduction

Currently, there are numerous methods for synthesizing composites [1—3 and other]. One of them is based on combustion synthesis or SHS [4]. The chemical reactions in powder mixtures are diverse. The obtained multi-component and multi-phase reaction products, depending on the conditions of synthesis, can possess different properties. In general case, the deformation and stress field formation processes accompanying the chemical transformations are inherent stages of the synthesis and may impact the effective properties and structure. Such mathematical models, taking into account the processes of different physical nature, are called coupled models. From this perspective, the models of classic combustion theory—taking into account the effect of heat liberation from the reactions on the temperature field—also belong to coupled models. Taking into account the mutual influence of thermal phenomena, chemical reactions and deformation, in the elementary case, affects the effective formal-kinetic parameters [5, 6]. In the case of more detailed investigation, it leads to the occurrence of new transformation regimes [7]. However, before the detailed accounting of mechanical and mechanochemical phenomena, the sequence and mutual influence of the chemical stages should be elucidated. This work is aimed at studying the regimes of transformation using a non-stationary model of composite synthesis that includes consecutive-parallel stages at the moment of powder mixture reaction initiation by a heat pulse applied from a side end.

*anna-knyazeva@mail.ru
†bnv@ispms.tsc.ru

1. Mathematical statement of problem

Let us assume that the composite synthesis can be described by a simple scheme of chemical reactions. The result of the first reaction are reinforcing particles and an intermediate product consumed in the second reaction. The second reaction immediately forms the matrix. The general reaction scheme can be expressed as

$$\begin{aligned} 2X + Y &\to P_1 + 2P_2, \\ X + 2P_2 + Z &\to P. \end{aligned} \tag{1}$$

This scheme can be exemplified by the synthesis of composites in systems $Al+Cr_2O_3+Ti$ and $Al+Fe_2O_3+Ni$. According to scheme (1), the reagent Y (oxides) takes part in the first reaction, consequently changing its rate, while the concentration of the component Z (titanium or nickel) in the initial mixture changes the rate of the second reaction. Therefore, these substances may not be included into the kinetic equations. The resulting set of kinetic equations, corresponding to the said reaction scheme, is as follows:

$$\begin{aligned} \frac{dP_1}{dt} &= k_1(T)\bar{X}^2, \\ \frac{dP_2}{dt} &= 2k_1(T)\bar{X}^2 - 2k_2(T){P_2}^2\bar{X}, \\ \frac{dP_1}{dt} &= k_2(T){P_2}^2\bar{X}, \end{aligned} \tag{2}$$

where $\bar{X} = 1 - P - P_1 - P_2$ is the total number of reagents, $k_1(T) = k_{10}\exp\left(-\dfrac{E_1}{RT}\right)$ and $k_2(T) = k_{20}\exp\left(-\dfrac{E_2}{RT}\right)$ are the rates of reactions 1 and 2. Assuming that the synthesis is initiated from the surface by a uniformly distributed heat source (heat pulse) and the side surfaces are heat-insulated, in the thermal section of the problem we will stick to the one-dimensional thermal conductivity equation with chemical sources:

$$c\rho\frac{\partial T}{\partial t} = \lambda\frac{\partial^2 T}{\partial x^2} + Q_1\Phi_1 + Q_2\Phi_2, \tag{3}$$

where $\Phi_1 = k_1(T)\bar{X}^2$, $\Phi_2 = k_2(T)P_2^2\bar{X}$.

The boundary conditions are:

$$x = 0 : -\lambda\frac{\partial T}{\partial x} = q_0,\ t < t_{imp}, \quad \text{and} \quad -\lambda\frac{\partial T}{\partial x} = 0, \quad t > t_{imp}, \tag{4}$$

$$x \to \infty : \frac{\partial T}{\partial x} = 0. \tag{5}$$

To reduce the number of variables and number of necessary numerical calculations, let us introduce the following dimensionless variables: $\theta = \dfrac{T - T_*}{T_* - T_0}$, $\tau = \dfrac{t}{t_*}$, $\xi = \dfrac{x}{x_*}$, where $t_* = \dfrac{c\rho RT_*^2}{k_{10}E_1Q_1}\exp\left(\dfrac{E_1}{RT_*}\right)$, $T_* = \dfrac{Q_1}{c\rho} + T_0$ and $x_*^2 = \dfrac{\lambda t_*}{c\rho}$ are the characteristic scales.

Then, the problem (2)–(5) will take the following form:

$$\frac{\partial\theta}{\partial\tau} = \frac{\partial^2\theta}{\partial\xi^2} + W_{ch}, \tag{6}$$

$$\frac{\partial P_1}{\partial\tau} = \bar{X}^2\gamma\exp\left(\frac{\theta\sigma}{\beta(1+\theta\sigma)}\right), \tag{7}$$

$$\frac{\partial P_2}{\partial \tau} = 2\bar{X}\gamma\left[\bar{X}\exp\left(\frac{\theta\sigma}{\beta(1+\theta\sigma)}\right) - zP_2^2\exp\left(\frac{\theta\sigma+\varepsilon}{\beta(1+\theta\sigma)}\right)\right], \tag{8}$$

$$\frac{\partial P}{\partial \tau} = z\bar{X}P_2^2\gamma\exp\left(\frac{\theta\sigma+\varepsilon}{\beta(1+\theta\sigma)}\right), \tag{9}$$

$$\xi \to 0 : -\frac{\partial\theta}{\partial\xi} = Q_e, \quad \tau < \tau_* \quad \text{and} \quad -\frac{\partial\theta}{\partial\xi} = 0, \quad \tau > \tau_* \tag{10}$$

$$\xi \to \infty : \ \frac{\partial\theta}{\partial\xi} = 0, \tag{11}$$

$$\tau = 0 : \theta = -1, \quad P = P_1 = P_2 = 0,$$

where $\tau_* = \dfrac{t_{imp}}{t_*}$, $W_{ch} = \dfrac{\beta}{\sigma}\bar{X}\left(\bar{X}\exp\left(\dfrac{\theta\sigma}{\beta\,(1+\theta\sigma)}\right) + K_Q z P_2^2\exp\left(\dfrac{\theta\sigma+\varepsilon}{\beta\,(1+\theta\sigma)}\right)\right)$.

The problem, along with the pulse duration, includes dimensionless parameters $\sigma = \dfrac{T_* - T_0}{T_*}$, $\beta = \dfrac{RT_*}{E_1}$ is the lesser parameter of the combustion theory, $\gamma = \dfrac{c\rho RT_*^2}{E_1 Q_1} << 1$ is the lesser parameter of the combustion theory that characterizes the sensitivity of the reaction rate to burning-out, $\theta_0 = \dfrac{\sigma}{\beta}$ is the temperature head, while the parameters $K_Q = \dfrac{Q_2}{Q_1}$, $z = \dfrac{k_2}{k_1}$, $\varepsilon = \dfrac{E_1 - E_2}{E_1}$ characterize the relations between the kinetic reaction parameters, $Q_e = \dfrac{q_0}{Q_1}\sqrt{\dfrac{t_*}{\kappa}}$ is the relation of heat accumulated in the layer x_* to the reaction heat Q_1, $\kappa = \dfrac{\lambda}{c\rho}$.

2. Method of solution

The thermal conductivity equation (6) was solved using am implicit difference scheme of first-order approximation in time and second-order approximation in spatial coordinates and implementing the sweep method [8]. The kinetic equations for the total reactions were solved by an explicit-implicit method, which efficacy was shown elsewhere [9]. For instance, let us express eq. (7) for the numerical solution as

$$\frac{P_{1_i} - \check{P}_{1_i}}{\Delta\tau} = \gamma\left(1 - \check{P}_i - \check{P}_{1_i} - \check{P}_{2_i}\right)(1 - \check{P}_i - P_{1_i} - \check{P}_{2_i})\exp\left(\frac{\theta\sigma}{\beta(1+\theta\sigma)}\right)$$

or

$$P_{1_i} = \frac{\check{P}_{1_i} + \left(1 - \check{P}_i - \check{P}_{2_i}\right)Z}{1+Z}, \tag{12}$$

where $Z = \Delta\tau\gamma\left(1 - \check{P}_i - \check{P}_{1_i} - \check{P}_{2_i}\right)\exp\left(\dfrac{\theta\sigma}{\beta(1+\theta\sigma)}\right)$. In eq. (10), the numerator is, obviously, always less than the denominator. Similarly for eqs. (8) and (9):

$$P_{2_i} = \frac{\check{P}_{2_i} + \left(1 - \check{P}_i - \check{P}_{1_i}\right)Z}{1 + 2Z + 2Z_2}, \tag{13}$$

$$P_i = \frac{\check{P}_i + \left(1 - \check{P}_{1_i} - \check{P}_{2_i}\right)Z}{1 + Z_3}, \tag{14}$$

where $Z_2 = \Delta\tau\gamma z\check{P}_{2_i}\left(1 - \check{P}_i - \check{P}_{1_i} - \check{P}_{2_i}\right)\exp\left(\dfrac{\theta\sigma+\varepsilon}{\beta(1+\theta\sigma)}\right)$, $Z_3 = \Delta\tau\gamma z\check{P}_{2_i}^2\exp\left(\dfrac{\theta\sigma+\varepsilon}{\beta(1+\theta\sigma)}\right)$.

In eq. (12), the following notations were accepted: $\check{P}_{1_i}$ is the value P_1 in point i of the difference grid at the moment of time $j\Delta\tau$; P_{1_i} is the value P_1 in point i of the difference grid at the moment of time $(j+1)\Delta\tau$.

To confirm the correctness of the results, the problem was solved using two algorithms: (1) consecutive calculations at each of the layers in time with the verification of the convergence with decreasing step in time and number of decompositions of the calculation domain; (2) using the iteration at each of the layers in time. In the second case, the iterations are repeated until the result ceases to change (in the selected points) with a given accuracy:

$$\left[\sum_{(k)}\left(\tilde{\theta}_k - \theta k\right)^2\right]^{-1/2} \leqslant \varepsilon,$$

where $\tilde{\theta}_k$ is the solution for the temperature obtained in the previous iteration, θ_k is current calculation.

Tabs. 1 and 2 contain the data on the temperature at specific moments of time in point $\xi = 0$ for $h = 0.005$ and different steps in time $\Delta\tau$. Both the tables illustrate satisfactory convergence with decreasing step in time. Both the algorithms demonstrate satisfactory convergence also with increasing number of decompositions of the calculation domain (i.e. with decreasing spatial step), which is not shown in the tables.

Table 1. Temperature and conversion degree in point $\xi = 0$ for different time steps for the problem solution as per the first algorithm

	dt	t=0.5s.	t=1s.	t=2s.
θ	0.0025	3.9386	10.7376	67.9454
	0.005	3.9153	10.6336	65.6274
	0.01	3.8695	10.4418	62.4896
P	0.0025	$1.5441 \cdot 10^{-4}$	0.0697	0.4722
	0.005	$1.5252 \cdot 10^{-4}$	0.00684	0.4702
	0.01	$1.488 \cdot 10^{-4}$	0.00658	0.466
P_1	0.0025	0.02591	0.09022	0.4876
	0.005	0.02589	0.08982	0.4856
	0.01	0.02584	0.08906	0.4826
P_2	0.0025	0.0515	0.1664	0.037
	0.005	0.0514	0.1657	0.04
	0.0512	0.0513	0.1645	0.046

The iteration method converges, on average, on the 2–6th iteration, depending on the process stage: heating, reaction initiation, or process stabilization. The advantage of the iteration method is in the possibility to use in calculations quite large time steps with satisfactory convergence of the kinetic subproblem. The results below were generally obtained with the following parameters of the problem: $\sigma = 0.9$, $\varepsilon = 0.01$, $\gamma = 0.035$, $K_Q = 0.3$, $z = 5$.

3. Results and discussion

A typical distribution of the temperatures and concentrations in the surface layer at different moments of time is shown in Fig. 1 for different external pulse durations. The small duration

Table 2. Temperature and conversion degree in point $\xi = 0$ for different time steps using iterations

	dt	t=0.5s.	t=1s.	t=2s.
θ	0.1	4.0828	12.3788	43.2305
	0.04	4.2018	11.3522	51.7492
	0.025	3.9876	11.1474	55.5466
P	0.1	0.00015	0.0111	0.4089
	0.04	0.00018	0.0084	0.4462
	0.025	0.0015	0.0078	0.4566
P_1	0.1	0.02912	0.1164	0.4655
	0.04	0.02879	0.0989	0.4744
	0.025	0.02666	0.0955	0.4776
P_2	0.1	0.05746	0.2015	0.0998
	0.04	0.05701	0.1786	0.0676
	0.025	0.05291	0.1740	0.0574

of the external pulse is insufficient for the reaction in the mixture to continue. Disabling the source inhibits the reaction, after which the reaction stops (Fig. 1, left side). With increased pulse duration, there is a self-sustained reaction and stabilization of the process (Fig. 1, right side). We suppose that the stationary regime was caused by the conditions when the maximum temperature does not change with the accuracy of 3–5%. Other criteria–connected with the analysis of various thermodynamic characteristics are also possible. The figures in the center correspond to some intermediate case, when the stabilization of the process takes long time. In this case, there are fluctuations in the composition at the initial stage of the process evolution.

The maximum temperature, evidently, grows with the increase of the pulse duration. However, during the stabilization process, the maximum temperature decreases and approaches some asymptotic limit. The same is applicable to the fraction of particles in the reaction products P_1 and the fraction of the matrix P. The intermediate product exists in a considerable concentration in the reaction front; however, then, the concentration P_2 appreciably drops due to the end product formation. In the combustion theory, the stationary regime is a limit that should not depend on the initiation conditions and properties of a system under study. In the investigation of the stabilization process, it cannot be shown. Having a short pulse, the reactions in the surface layer stop (Fig. 2b, solid lines). In the stationary regime, they condition some end composition of the composite that depends on the initial composition of the reagents and other parameters. In out model, the variation of the initial composition is bound with altered relation of pre-exponential factors of the reactions.

Fig. 3 demonstrate changed composition of the products. The alteration of parameter is connected with the change to the stationary velocity of the front that can be defined using various methods (though, all the methods are equivalent to each other). For instance, the velocity can be calculated from Fig. 3 using the data on the location of the point with fixed value of the product P concentration at different moments of time. For three values of $z = 0.5$, 3 and 5, we get $V = 0.365$, 0.592 and 0.735, respectively.

After changing parameters γ or β, we get the stationary regime of the composite synthesis with different velocity and different temperature in the front. However, the composition of the products remains unchanged.

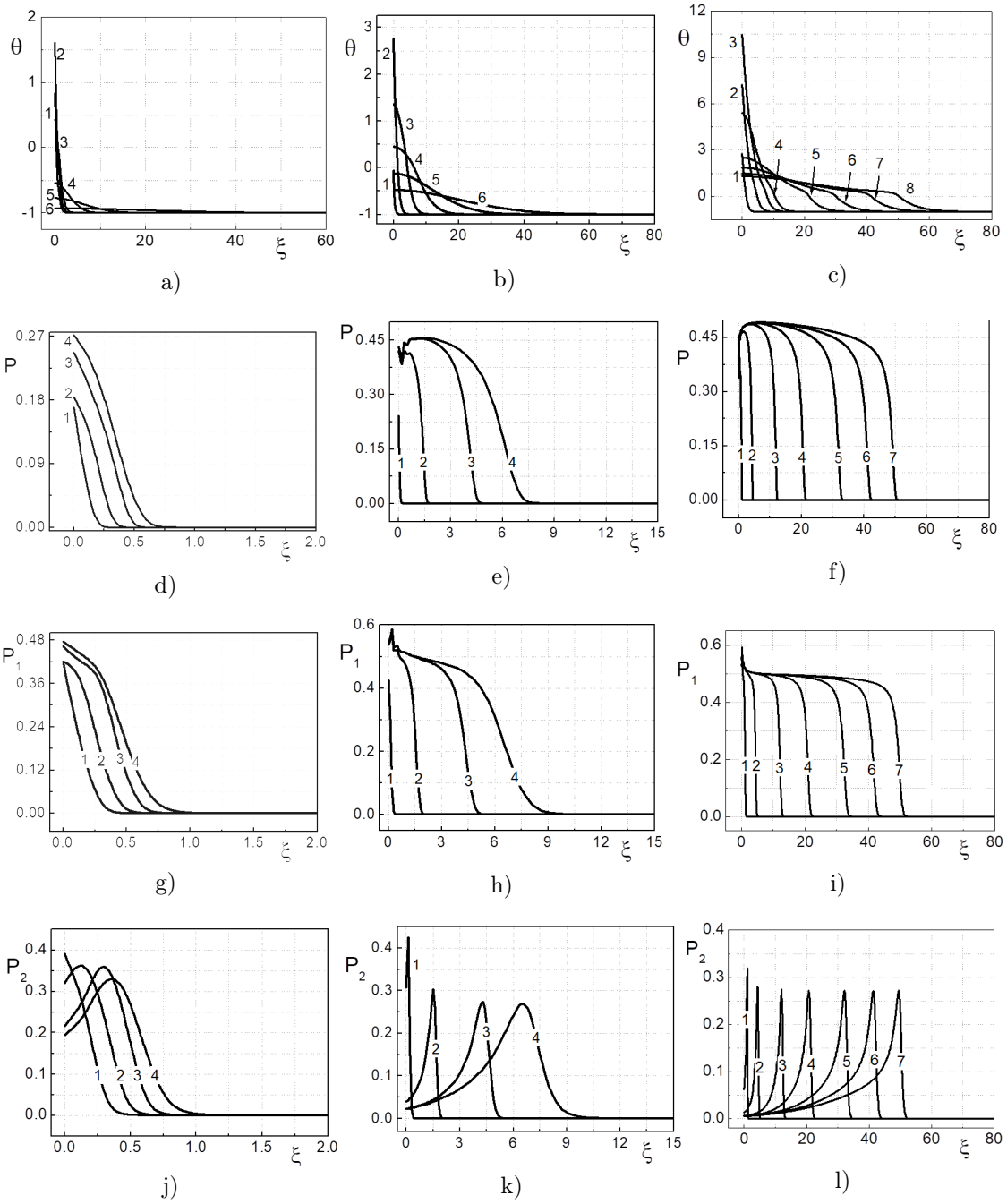
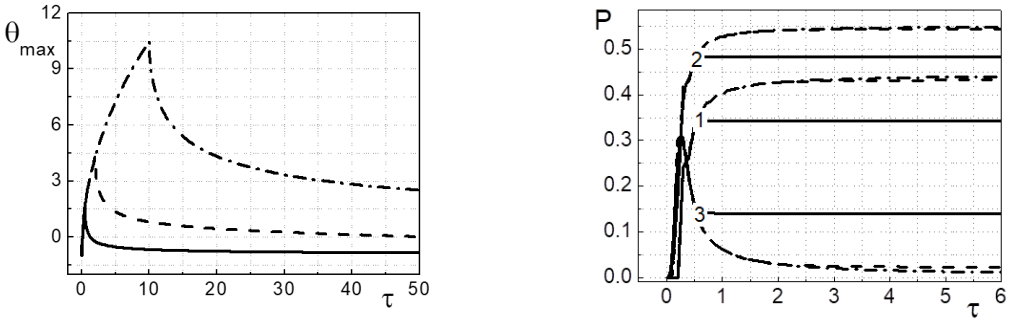
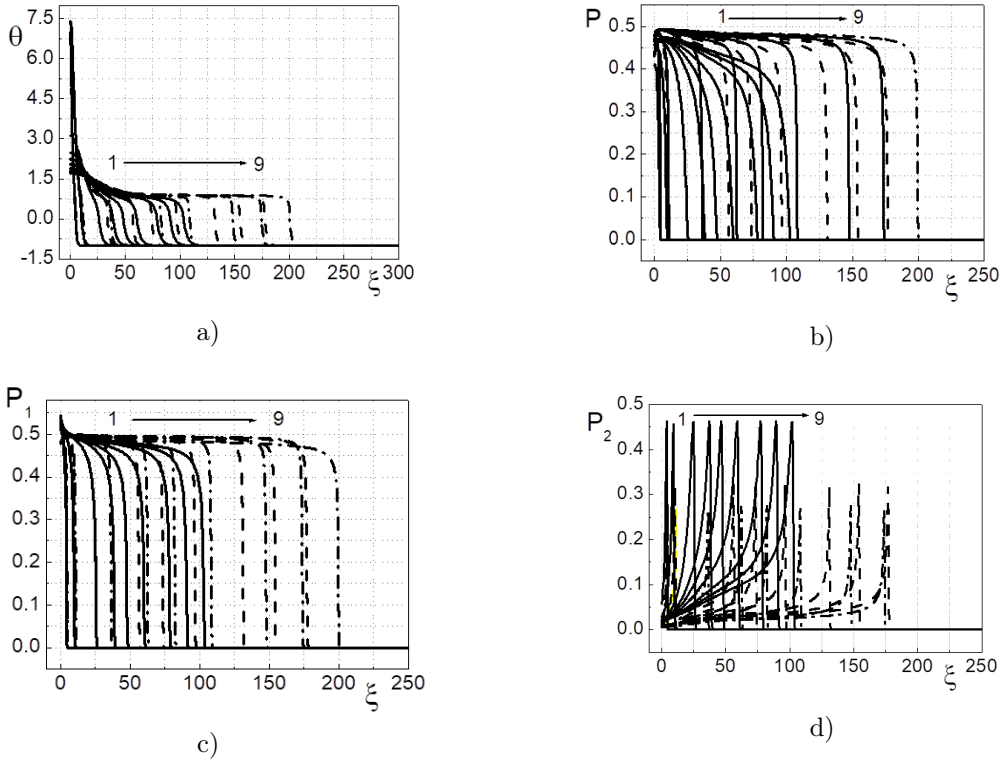


Fig. 1. Temperature and concentration distribution for different pulse durations τ_* at different moments of time. a), d), g) and j) correspond to $\tau_* = 0.5$; b), e), h) and k) correspond to $\tau_* = 2$; c), f), i) and l) correspond to $\tau_* = 10$. Moments of time: a) t = 1) 0.3, 2) 0.5, 3) 0.8, 4) 5, 5) 20, 6) 200; b) t = 1) 0.3, 2) 2, 3) 2.5, 4) 5, 5) 50, 6) 200; c) t = 1) 1, 2) 5, 3) 10, 4) 15, 5) 50, 6) 90, 7) 150, 8) 200; d, g, j) t = 1) 0.3, 2) 0.4, 3) 0.5, 4) 200; e, h, k) t = 1) 0.3, 2) 1.5, 3) 7, 4) 200; f, i, l) t = 1) 1, 2) 5, 3) 20, 4) 50, 5) 105, 6) 155, 7) 200



a) b)

Fig. 2. Time dependence of the maximum temperature and concentrations in the surface layer. The solid line corresponds to the calculation of the pulse duration $\tau_* = 0.5$; the dashed line corresponds to $\tau_* = 2$; the dash-dotted line corresponds to $\tau_* = 10$. Curves 1 in b) correspond to product P, the curves 2 correspond to product P_1, the curves 3 correspond to product P_2



a) b) c) d)

Fig. 3. Temperature and concentration distribution for different values of parameter z at different moments of time. The solid line corresponds to the parameter value of $z = 0.5$, the dashed line corresponds to $z = 3$, the dash-dotted line corresponds to $z = 5$, $\gamma = 0.025$. Moments of time: t = 1) 4, 2) 12, 3) 50, 4) 90, 5) 120, 6) 160, 7) 220, 8) 260, 9) 300

4. Transition to the coupled model

The mathematical model that takes into account the mutual influence of thermal, chemical and mechanical phenomena for the described reaction scheme and built similarly to [10, 11]

includes an equation that is similar to eq. (3), but including an additional nonchemical source of heat:

$$c'_\varepsilon \rho \frac{dT}{dt} = \lambda \frac{\partial^2 T}{\partial x^2} - W(t,x) + Q'_1 \Phi_1 + Q'_2 \Phi_2$$

where

$$c'_\varepsilon = c_\varepsilon \left(1 + 3\frac{K\alpha_T^2}{\rho c_\varepsilon} T \frac{1+\nu}{1-\nu}\right),$$

$$Q'_1 = Q_1 - 3K\alpha_T T \frac{1+\nu}{1-\nu}\left[(\alpha_1 - \alpha_x) + 2(\alpha_2 - \alpha_x)\right],$$

$$Q'_2 = Q_2 - 3K\alpha_T T \frac{1+\nu}{1-\nu}\left[(\alpha - \alpha_x) - 2(\alpha_2 - \alpha_x)\right],$$

$$W(t,x) = 2K\alpha_T \frac{2-4\nu}{1-\nu}\left(\frac{x}{h} - 1\right) T \frac{d}{dt} F(t,x),$$

$$F(t,x) = \frac{2}{h^2}\int_0^h w(t,x)x\,dx - \frac{4}{3h}\int_0^h w(t,x)\,dx,$$

$$w = 3\left[\alpha_T (T - T_0) + (\alpha_1 - \alpha_x)(P_2 - P_{2_0}) + (\alpha - \alpha_x)(P - P_0)\right],$$

K is the isothermal volume elasticity modulus, ν is the Poisson's ratio; α_T is the linearly coefficient of thermal expansion (averaged in the properties of the reagents and products); α, α_1, α_2,α_x are the coefficients of concentration expansion of the reagents and reaction products; h is the sample dimensions along 0X. The alteration of the effective properties (thermal capacity and thermal effects of the reaction), compared to the uncoupled model, can be interpreted as the expansion of the variation domain of the model parameters (6)–(11). However, the introduction of the additional heat source may introduce its own peculiarities; for instance, lead to the expansion of the domain of existence of stationary conversion regime, as compared to the classic works [12–15 and other] even with due account for the peculiarities in the kinetic functions that reflect the specificity of the reactions in heterogeneous systems. This should be investigated further.

Conclusions

Therefore, the work proposed the model of composite synthesis with reinforcing inclusions. The synthesis of the matrix and inclusions is determined by two total parallel-consecutive stages. Two algorithms for the problem solution on the reaction initiation give similar results. It was shown that in the model, there is a domain of parameters, where a stationary synthesis regime is possible. The composition of the products depends on the relation of the model parameters, including the changes to the initial composition, which in the model is connected with the alteration of the relation of the reaction pre-exponential factors. The paper presented a coupled model of the composite synthesis, including the effective properties (depending on the mechanical characteristics) and an additional heat source, conditioned by the interaction of different physical processes. The role of the new factors requires additional investigation.

The reported study was funded by RFBR, project no. 20-03-00303.

References

[1] S.-G.Son, B.-H.Lee, J.-M.Lee etc., Low-temperature synthesis of $(TiC + Al_2O_3)/Al$ alloy composites based on dopant-assisted combustion, *J. Alloys Compd.*, **649**(2015), 409–416.

[2] Y.M.Z.Ahmed, Z.I.Zakiab, R.K.Bordiac etc., Simultaneous synthesis and sintering of TiC/Al_2O_3 composite via self propagating synthesis with direct consolidation technique, *Ceram. Int.*, **42**(2016), 16589–16597.

[3] D.Horvitz, I.Gotman, E.Y.Gutmanas, etc., In situ processing of dense $Al_2O_3^{\smile}Ti$ aluminide interpenetrating phase composites, *J. Eur. Ceram. Soc.*, **22**(2002), 947–954.

[4] A.S.Mukasyan, A.S.Rogachev, Combustion synthesis: mechanically induced nanostructured materials, *J. Mater. Sci.*, **52**(2017), 11826–11833. DOI: 10.1007/s10853-017-1075-9

[5] S.N.Sorokova, A.G.Knyazeva, Numerical study of the influence of the technological parameters on the composition and stressed-deformed state of a coating synthesized under electron-beam heating, *Theor. Found. Chem. Engi.*, **44**(2010),no. 2, 172–185.
DOI: 10.1134/S0040579510020089

[6] A.G.Knyazeva, Velocity of the simplest solid-phase chemical reaction front and internal mechanical stresses, *Combust. Explos. Shock Waves*, **30**(1994),no.1, 43–53.

[7] A.G.Knyazeva, Solution of the Thermoelasticity Problem in the Form of a Traveling Wave and its Application to Analysis of Possible Regimes of Solid?Phase Transformations, *J. Appl. Mech. Tech. Phys.*, **44**(2003), no. 2, 164–173. DOI: 10.1007/s10573-006-0087-6

[8] V.M.Paskonov, V.I.Polezhaev, L.A.Chudov, Numerical modeling of the heat and mass exchange processes, Moscow, Nauka, 1984 (in Russian).

[9] V.N.Demidov, A.G.Knyazeva, Multistage kinetics of the synthesis $Ti - T_xC_y$ composite, *Nanoscience and Technology: An International Journal*, **10**(2019), no. 3, 195–218.

[10] A.G.Knyazeva, Effect of fixing conditions on the heating rate of a specimen, *Combust. Explos. Shock Waves*, **36**(2000), no. 5, 582–590.

[11] A.G.Knyazeva, O.N.Kryukova, The propagation of solid-phase transformation in a flat layer with due regard to the coupling of thermal and mechanical processes, *Math. modeling*, **15**(2003), no. 8, 21–33 (in Russian).

[12] A.P.Aldushin, Stationary propagation of an exothermic-reaction front in a condensed medium, *J. Appl. Mech. Tech. Phys.*, **15**(1974), no. 3, 96–105.

[13] M.B.Borovikov, I.A.Burovoi, U.I.Gol'dshleger, Combustion wave propagation in systems of sequential reactions with endothermal stages, *Combust. Explos. Shock Waves*, **20**(1984), no. 3, 241–248.

[14] E.A.Nekrasov, A.M.Timokhin, Theory of thermal wave propagation of multistage reactions described by simple empirical schemes, *Combust. Explos. Shock Waves*, **22**(1986), no. 4, 431–437.

[15] A.P.Aldushin, T.M.Martem'yanova, A.G.Merzhanov ets., Propagation of the front of an exothermic reaction in condensed mixtures with the interaction of the components through a layer of high-melting product, *Combust. Explos. Shock Waves*, **8**(1972), no. 2, 202–212.

Связанная математическая модель синтеза композитов

Анна Г. Князева
Наталья В. Букрина
Институт физики прочности и материаловедения СО РАН
Томск, Российская Федерация

Аннотация. В работе предложена модель синтеза композита «металлическая матрица – упрочняющие включения». Решение осуществлено с помощью двух алгоритмов, показывающих близкие результаты. Показано, что, как и в классических моделях горения, существует область параметров модели, в которой возможен переход к стационарному режиму. Продемонстрировано, что учет связанности тепловых и химических процессов приводит к изменению эффективных свойств (теплоемкости и тепловых эффектов реакции) и появлению дополнительного источника тепла, обусловленного взаимодействием процессов разной физической природы.

Ключевые слова: синтез композитов, импульсный нагрев, последовательно-параллельные реакции, сравнение алгоритмов.

DOI: 10.17516/1997-1397-2020-13-6-718-732
УДК 517.55; 517.9

Upper Bounds for the Analytic Complexity of Puiseux Polynomial Solutions to Bivariate Hypergeometric Systems

Vitaly A. Krasikov*
Plekhanov Russian University of Economics
Moscow, Russian Federation

Received 10.06.2020, received in revised form 24.07.2020, accepted 20.09.2020

Abstract. The paper deals with the analytic complexity of solutions to bivariate holonomic hypergeometric systems of the Horn type. We obtain estimates on the analytic complexity of Puiseux polynomial solutions to the hypergeometric systems defined by zonotopes. We also propose algorithms of the analytic complexity estimation for polynomials.

Keywords: hypergeometric systems of partial differential equations, holonomic rank, polynomial solutions, zonotopes, analytic complexity, differential polynomial, hypergeometry package.

Citation: V.A. Krasikov,Upper Bounds for the Analytic Complexity of Puiseux Polynomial Solutions to Bivariate Hypergeometric Systems , J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 718–732.
DOI: 10.17516/1997-1397-2020-13-6-718-732.

1. Introduction and preliminaries

The notion of complexity is widely used in Mathematics and Computer Science in the context of several various abstract objects. The computational complexity of algorithms, the algebraic complexity of polynomials, the Rademacher complexity in the computational learning theory or the social complexity in the social systems are the concepts of great importance in the corresponding fields of science. The present work is devoted to the particular type of complexity – the analytic complexity of bivariate holomorphic functions.

The notion of analytic complexity is closely related to Hilbert's 13th problem, which was solved by A. N. Kolmogorov and V. I. Arnold in 1957 [1]. The initial formulation of Hilbert's 13th problem asks whether any continuous function of several variables can be represented as a finite superposition of bivariate functions [17]. The problem of finding similar representations for analytic functions has given rise to the theory of the analytic complexity. The main objects under consideration in this theory are the *analytic complexity classes.*

Definition 1 (See [2]). *Let $\mathcal{O}(U(x_0, y_0))$ denote the set of holomorphic functions in an open neighborhood $U(x_0, y_0)$ of a point $(x_0, y_0) \in \mathbb{C}^2$. The class Cl_0 of analytic functions of analytic complexity zero is defined to comprise the functions that depend on at most one of the variables. A function $f(x, y)$ is said to belong to the class Cl_n of functions with analytic complexity $n > 0$ if there exists a point $(x_0, y_0) \in \mathbb{C}^2$ and a germ $\mathfrak{f}(x, y) \in \mathcal{O}(U(x_0, y_0))$ of this function holomorphic at (x_0, y_0) such that $\mathfrak{f}(x, y) = c(a(x, y) + b(x, y))$ for some germs of holomorphic functions $a, b \in Cl_{n-1}$ and $c \in Cl_0$. If there is no such representation for any finite n, then the function f is said to be of infinite analytic complexity.*

*Krasikov.VA@rea.ru

Example 1. A generic element of the first complexity class Cl_1 is a function of the form $f_3(f_1(x) + f_2(y))$. A function in Cl_2 can be represented in the form $f_7\left(f_5(f_1(x) + f_2(y)) + f_6(f_3(x) + f_4(y))\right)$, where $f_i(\cdot)$ are univariate holomorphic functions, $i = 1, \ldots, 7$.

For any class of analytic complexity $Cl_n, n \in \mathbb{N}$ there exists a system of differential polynomials with constant coefficients Δ_n which annihilates a function if and only if it belongs to Cl_n.

Example 2 (See [2])**.** For a bivariate function $f(x, y)$ consider the differential polynomial

$$\Delta_1(f) = f'_x(f'_y)^2 f'''_{xxy} - (f'_x)^2 f'_y f'''_{xyy} + f''_{xy}(f'_x)^2 f''_{yy} - f''_{xy}(f'_y)^2 f''_{xx}.$$

This differential polynomial vanishes if and only if its argument $f \in Cl_1$.

The problem of defining whether a function belongs to an analytic complexity class is equivalent to computing the corresponding system of differential polynomials. Note that this is a problem of formidable computational complexity [4, 11] and a direct approach to its solution appears to be inappropriate.

An important question is a possible connection between the classes of finite analytic complexity and hypergeometric functions. In this paper we consider hypergeometric functions as solutions of hypergeometric systems in the sense of Horn [8, 10]. We choose a matrix $A \in \mathbb{Z}^{m\times n} = (A_{ij},\ i = 1, \ldots, m, j = 1, \ldots, n)$ and a vector of parameters $c = (c_1, \ldots, c_m) \in \mathbb{C}^m$. We denote the rows of this matrix by $\mathbf{A}_i$, $i = 1, \ldots, m$.

Definition 2. *The hypergeometric system (or the Horn system) Horn(A, c) is the following system of partial differential equations:*

$$x_j P_j(\theta) f(x) = Q_j(\theta) f(x),\ j = 1, \ldots, n, \tag{1}$$

where

$$P_j(s) = \prod_{i:A_{ij}>0} \prod_{l_j^{(i)}=0}^{A_{ij}-1} \left(\langle \boldsymbol{A}_i, s\rangle + c_i + l_j^{(i)}\right),$$

$$Q_j(s) = \prod_{i:A_{ij}<0} \prod_{l_j^{(i)}=0}^{|A_{ij}|-1} \left(\langle \boldsymbol{A}_i, s\rangle + c_i + l_j^{(i)}\right),$$

and $\theta = (\theta_1, \ldots, \theta_n)$, $\theta_j = x_j \dfrac{\partial}{\partial x_j}$.

It has been conjectured in [14] that any hypergeometric function has finite analytic complexity. Hypergeometric systems of equations differ greatly from the differential criteria for the analytic complexity classes, but numerous computer experiments suggest that the conjecture is true in a lot of particular cases [6, 7]. The case of hypergeometric systems with low holonomic rank has been considered in [9].

The set of functions of infinite analytic complexity is also a matter of interest. Until recently, all known examples of such functions were differentially transcendental functions, that is, functions that are not solutions to any nonzero differential polynomial with constant coefficients. Important examples of differentially algebraic functions of infinite analytic complexity have been presented in [15, 16].

A bivariate hypergeometric system can be defined by an integer convex polygon and a complex vector of parameters as explained in the next definition.

Definition 3. *Let l_i denote the generator of the sublattice $\{s \in \mathbb{Z}^n : \langle \boldsymbol{A}_i, s\rangle = 0\}$ and let k_i be the number of elements in the set $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m\}$, which coincide with $\boldsymbol{A}_i$. Let us define the polygon $\mathcal{P}(A)$ (see [13]) as the integer convex polygon whose sides are translations of the vectors $k_i l_i$, the vectors $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ being the outer normals to its sides. We will say that the hypergeometric system Horn(A, c) is defined by the polygon $\mathcal{P}(A)$ and the vector $c \in \mathbb{C}$.*

Definition 4. *A polygon is called a zonotope if it can be represented as the Minkowski sum of segments.*

In this article we investigate the analytic complexity of solutions to hypergeometric systems of equations (1) defined by zonotopes.

The present paper is organized as follows. In Section 2 we investigate particular cases of hypergeometric systems defined by zonotopes and analyze the analytic complexity of their solutions. We formulate and prove an estimate of the analytic complexity for Puiseux polynomial solutions to such systems in terms of the defining matrices and parameter vectors. In Section 3 we present algorithms for finding the supports of Puiseux polynomial solutions to hypergeometric systems and estimating the analytic complexity of polynomials. In Section 4 we consider examples of hypergeometric systems and estimate the analytic complexity of their solutions. Throughout the rest of the paper by «polynomial solutions to hypergeometric systems» we mean Puiseux polynomial solutions.

We use the Wolfram Mathematica package HyperGeometry for solving hypergeometric systems we investigate in this article. The package is available for free public use at https://www.researchgate.net/publication/318986894_HyperGeometry, the description of available functions is given in [12].

2. Hypergeometric systems defined by zonotopes

Let us consider the special case of hypergeometric systems defined by zonotopes. Numerous experiments suggest that the analytic complexity of polynomial solutions to such systems can be much lower than its estimate based on their supports (see [3, Proposition 4]).

The set of hypergeometric systems defined by zonotopes enjoys the following properties:

a) these systems are holonomic for generic values of parameters;

b) the holonomic rank of a hypergeometric system (see Theorem 2.5 in [5]) is given by

$$\operatorname{rank}(\operatorname{Horn}(A, c)) = d_1 d_2 - \sum_{\mathbf{A}_i, \mathbf{A}_j \text{ lin. dependent}} \nu_{ij}, \tag{2}$$

where $d_j = \sum\limits_{\substack{i=1 \\ A_{ij} > 0}}^{m} A_{ij}, j = 1, 2$ and

$$\nu_{ij} = \begin{cases} \min(|A_{i1}A_{j2}|, |A_{j1}A_{i2}|), \text{ if } \mathbf{A}_i, \mathbf{A}_j \text{ are in opposite open quadrants of } \mathbb{Z}^2, \\ 0, \text{ otherwise.} \end{cases}$$

For the hypergeometric systems defined by zonotopes there is another formula for computing their holonomic rank (see Proposition 1 in [9]), which in some cases may be more suitable;

c) for any number of rows (a_i, b_i) belonging to the matrix A defining such a system, A contains the same number of rows $(-a_i, -b_i)$. Thus the rows of A can be grouped into two matrices $\hat{A}, -\hat{A}$. This representation is in general not unique.

d) for a hypergeometric system defined by a zonotope one can always choose parameter values such that any solution to the resulting system is a polynomial (see [10, Proposition 6.5]). Namely, for such a hypergeometric system $\mathrm{Horn}(A, c)$, where the matrix A contains $2k$ rows, let $\alpha = (\alpha_1, \ldots, \alpha_k)$ be the part of the parameter vector c, corresponding to the matrix $\hat{A}$ (see the property (c) above), $\beta = (\beta_1, \ldots, \beta_k)$ be the part of this vector corresponding to $-\hat{A}$. By Proposition 4.7 in [10] the general solution to $\mathrm{Horn}(A, c)$ is a polynomial if $-\alpha_i - \beta_i \in \mathbb{N}\backslash\{0\}$ for $i = 1, \ldots, k$.

The simplest instance of a zonotope is a parallelogram. The analytic complexity estimate of the solutions to the systems defined by parallelograms is the basis for more complex cases.

Proposition 1. *The analytic complexity of a solution to a hypergeometric system defined by a parallelogram cannot exceed* 2.

Proof. The solutions to the hypergeometric system $\mathrm{Horn}(A, c)$ defined by a parallelogram have been described in Proposition 4.7 in [10]. For a bivariate system ($n = 2$) this formula leads to

$$(x_1^{-a_{11}} x_2^{-a_{21}})^{\alpha_1} \left(1 + x_1^{-a_{11}} x_2^{-a_{21}}\right)^{-\alpha_1-\beta_1} \cdot (x_1^{-a_{12}} x_2^{-a_{22}})^{\alpha_2} \left(1 + x_1^{-a_{12}} x_2^{-a_{22}}\right)^{-\alpha_2-\beta_2},$$

where $A^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, c = (\alpha_1, \alpha_2, \beta_1, \beta_2)$. The monomials $x_1^{-a_{11}} x_2^{-a_{21}}$ and $x_1^{-a_{12}} x_2^{-a_{22}}$ both belong to Cl_1, thus for any univariate analytic functions $\phi(\cdot), \psi(\cdot)$ the product $\phi(x_1^{-a_{11}} x_2^{-a_{21}}) \cdot \psi(x_1^{-a_{12}} x_2^{-a_{22}})$ belongs to Cl_2. □

The following example shows that the solutions to hypergeometric systems defined by more complex polygons can still have low analytic complexity.

Example 3. *A simple zonotope.* Let us consider the hypergeometric system $\mathrm{Horn}(A', c')$ defined by the matrix $A' = \begin{pmatrix} 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & -1 \end{pmatrix}^{\mathrm{T}}$ and the parameter vector $c' = (-23, 22, -10, 0, -9, 0)$. Using the formula (2) we conclude that the holonomic rank of this system is equal to 3. The hypergeometric system $\mathrm{Horn}(A', c')$ is defined by the zonotope shown in Fig. 1.

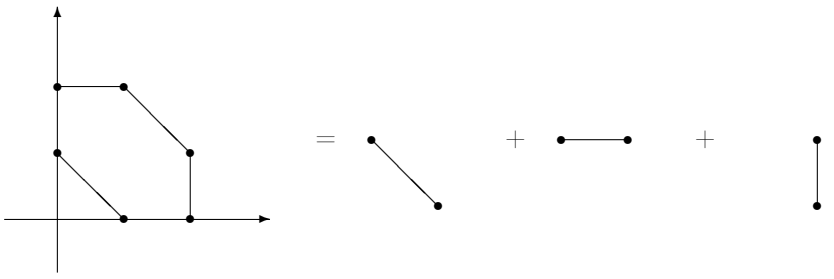


Fig. 1. Polygon defining the system $\mathrm{Horn}(A', c')$, and its representation as the Minkowski sum of segments

The support of the polynomial solutions to the system $\mathrm{Horn}(A', c')$ is shown in Fig. 2.

Let us consider the part of the solution $p_0(x, y)$ whose support is bounded by the straight lines parallel to the coordinate axes. Note that $p_0(x, y)$ contains 110 monomials (we do not put here the whole expression due to its large size) and the known estimates for polynomials [3, Proposition 4] imply that the analytic complexity of $p_0(x, y)$ does not exceed 5. Indeed, the support of $p_0(x, y)$ lies in the union of 10 lines parallel to the s axis. The analytic complexity of the polynomial

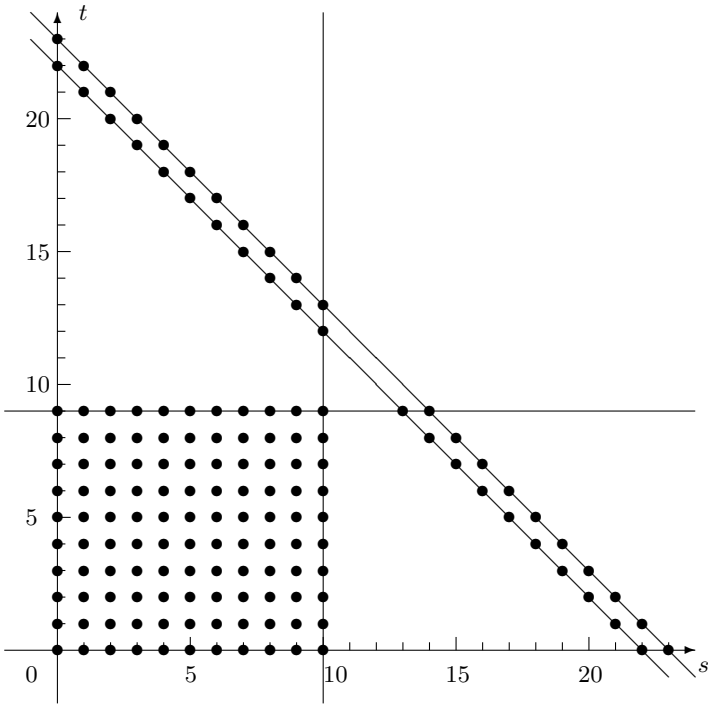


Fig. 2. The support for the solution of the system $\mathrm{Horn}(A', c')$

whose support lies on a straight line parallel to an axis cannot exceed 1. Then the analytic complexity of the sum of k such polynomials cannot exceed $1 + \lceil \log_2 k \rceil$, where by $\lceil x \rceil, x \in \mathbb{R}$ we denote the smallest integer not exceeding x. Later we prove that in fact the analytic complexity of $p_0(x, y)$ does not exceed 3.

In general, appending a pair of rows $(a_i, b_i), (-a_i, -b_i)$ to the matrix defining a hypergeometric system is equivalent to adding a pair of parallel straight lines bounding the support of the solution in the exponent space. Let the hypergeometric system be defined by a parallelogram, and let $p_0(x, y) = \sum\limits_{(s,t)\in S} c_{s,t} \cdot x^s y^t$ be a polynomial solution of this system with the support S. Adding a pair of straight lines in the exponent space leads to the system whose solution is given by

$$p_1(x, y) = \sum_{(s,t)\in S} \frac{\Gamma(\alpha_1 s + \beta_1 t + \gamma_1 + 1)}{\Gamma(\alpha_1 s + \beta_1 t + \gamma_1)} \cdot c_{s,t} \cdot x^s y^t = \sum_{(s,t)\in S} (\alpha_1 s + \beta_1 t + \gamma_1)\, x^s y^t =$$

$$= (\alpha_1 \theta_x + \beta_1 \theta_y + \gamma_1) \sum_{(s,t)\in S} c_{s,t} x^s y^t = (\alpha_1 \theta_x + \beta_1 \theta_y + \gamma_1)\, p_0(x, y).$$

Using this formula repetitively we obtain the solution for k additional pairs of rows $(a_i, b_i), (-a_i, -b_i)$:

$$p_k(x, y) = \left(\prod_{j=1}^{k} (\alpha_j \theta_x + \beta_j \theta_y + \gamma_j) \right) p_0(x, y).$$

Thus the estimate for the analytic complexity of $p_k(x, y)$ depends on the analytic complexity of $p_0(x, y)$. This dependence is described in detail in the following Proposition and its corollaries. Recall that we use the notation $\theta_x = x\dfrac{\partial}{\partial x}$, $\theta_y = y\dfrac{\partial}{\partial y}$ and $\alpha_j, \beta_j, \gamma_j \in \mathbb{C}$, $j = 1, \ldots, k$.

Proposition 2. *If $f(x,y) \in Cl_n$ then $(\alpha\theta_x + \beta\theta_y + \gamma)f(x,y) \in Cl_{2n+1}$.*

Proof. We use induction by n to show that $(\alpha\theta_x + \beta\theta_y)f(x,y) \in Cl_{2n}$.

For $n = 1$ we can represent $f(x,y)$ in the form $f(x,y) = c(a(x) + b(y))$.

$$(\alpha\theta_x + \beta\theta_y)c(a(x) + b(y)) = c'(a(x) + b(y)) \cdot (\alpha x a'(x) + \beta y b'(y)),$$

and this function belongs to Cl_2 as a product of Cl_1 functions. If the statement holds for all $n < N$, and $f(x,y)$ belongs to Cl_N, which means it can be represented as $f(x,y) = h(f_1(x,y) + f_2(x,y))$, where $f_1(x,y), f_2(x,y) \in Cl_{N-1}$, then

$$(\alpha\theta_x + \beta\theta_y)h(f_1(x,y) + f_2(x,y)) =$$

$$= h'(f_1(x,y) + f_2(x,y))\left((\alpha\theta_x + \beta\theta_y)f_1(x,y) + (\alpha\theta_x + \beta\theta_y)f_2(x,y)\right).$$

Both of the functions $f_1(x,y)$ and $f_2(x,y)$ belong to Cl_{N-1}, so the estimate of the analytic complexity for $(\alpha\theta_x + \beta\theta_y)f_i(x,y)$, $i = 1,2$ is Cl_{2N-2}. Then their sum belongs to Cl_{2N-1} and, after the multiplication of the result by $h'(f_1(x,y) + f_2(x,y)) \in Cl_N$, the product belongs to Cl_{2N}. Thus we conclude that for any n, if $f(x,y) \in Cl_n$ then $(\alpha\theta_x + \beta\theta_y)f(x,y) \in Cl_{2n}$. Adding $\gamma f(x,y) \in Cl_n$ to this expression we obtain a function in Cl_{2n+1}. □

Corollary 1. *For any $f(x,y) \in Cl_n$ the analytic complexity of*

$$\left(\prod_{j=1}^{k}(\alpha_j\theta_x + \beta_j\theta_y + \gamma_j)\right)f(x,y)$$

cannot exceed $2^k(n+1) - 1$.

Corollary 2. *Assume that the analytic complexity of a polynomial solution $p_0(x,y)$ to the hypergeometric system Horn$(\tilde{A}, \tilde{c})$ does not exceed n, S is a support of $p_0(x,y)$. Let the matrix A be obtained from $\tilde{A}$ by appending k pairs of vectors $(a_i, b_i), (-a_i, -b_i)$, vector c be obtained from $\tilde{c}$ by appending $2k$ elements. Then the analytic complexity of a polynomial solution with the support S to the hypergeometric system Horn(A, c) does not exceed $2^k(n+1) - 1$.*

Example 3. *(Continued).* Let us use Corollary 2 to estimate the analytic complexity of a solution to the system Horn(A', c'). To do this, consider the system Horn$(\tilde{A}', \tilde{c}')$, defined by the matrix $\tilde{A}' = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix}^{\mathrm{T}}$ and the vector of parameters $\tilde{c}' = (-10, -9, 1, 1)$. This system differs from the original one only by the absence of the pair of straight lines with the normal vectors $(1,1)$ and $(-1,-1)$ bounding the support of the solution. Thus this support for the system Horn$(\tilde{A}', \tilde{c}')$ coincides with the support of $p_0(x,y)$. Note that this system is defined by a parallelogram and hence by Proposition 2 the analytic complexity of its solutions cannot exceed 2. Computations show that the basis in the space of solutions to the system Horn$(\tilde{A}', \tilde{c}')$ consists of the single function $(x-1)^{10}(y-1)^9 \in Cl_1$, and hence $p_0(x,y) \in Cl_3$ by Corollary 2. The supports of two other solutions to Horn(A', c') lie on two parallel straight lines, so a linear combination of these solutions belongs to Cl_3, and the general solution to Horn(A', c') is a function in Cl_4.

The following theorem is the main theoretical result of the paper. It contains the general estimate of the analytic complexity for polynomial solutions to hypergeometric systems defined by zonotopes.

Theorem 1. *Let Horn(A,c) be a hypergeometric system defined by a zonotope. Assuming that the matrix A contains $2k$ rows, consider the matrices $\hat{A}$ and $-\hat{A}$ such that the union of their rows coincides with the set of rows of A. Let α be a part of the parameter vector c corresponding to the matrix $\hat{A}$, β be a part of this vector corresponding to $-\hat{A}$, and define the vector $\hat{c} = (\hat{c}_1, \dots, \hat{c}_k)$ by $\hat{c}_i = -\alpha_i - \beta_i$.*

If $\hat{c}_i \in \mathbb{N}\backslash\{0\}, i = 1, \dots, k$, then the analytic complexity of the general solution to Horn(A,c) does not exceed

$$\min\left(3\cdot 2^{k-2} - 1 + \left\lceil \log_2 \frac{k(k-1)}{2}\right\rceil, 2 + \lceil \log_2(\max_{i=1,\dots,k} \hat{c}_i + 1)\rceil + \lceil \log_2(k-1)\rceil\right).$$

Proof. For any system defined by a parallelogram the condition $\hat{c}_i \in \mathbb{N}\backslash\{0\}$ provides the existence of a polynomial solution (see [10, Proposition 4.7] and the proof of Proposition 2.). Appending of the rows $(a_i, b_i), (-a_i, -b_i)$ to the matrix defining the hypergeometric system affects only the coefficients of this solution but not its support. Without loss of generality we can choose a vector of parameters c such that the support of the general solution to Horn(A,c) coincides with a union of supports of the solutions to a finite number of systems defined by parallelograms (see proof of Proposition 6.5 in [10]). Thus the condition $\hat{c}_i \in \mathbb{N}\backslash\{0\}$ provides the existence of a polynomial basis in the space of solutions to Horn(A,c).

The matrix A contains $2k$ rows, so supports of the solutions are bounded by k pairs of straight lines. Let us assign a natural number from 1 to k to each pair of lines. The union of these supports is a subset of $\dfrac{k(k-1)}{2}$ parallelogram intersections (it is the sum of an arithmetic progression), each intersection we denote as $\square_{i,j}$, where $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, k\}$ are numbers assigned to pairs of straight lines which form the intersection, $i < j$. For any $(i,j) \in \{1, \dots, k\}^2$ the solution to Horn(A,c) whose support lies in the intersection $\square_{i,j}$ belongs to $Cl_{3\cdot 2^{k-2}-1}$ (by Corollary 2). The analytic complexity of the sum of $\dfrac{k(k-1)}{2}$ functions in $Cl_{3\cdot 2^{k-2}-1}$ (that is, the analytic complexity of the general solution to Horn(A,c)) cannot exceed $estim_1\left(\bigcup\limits_{i,j=1}^{k} \square_{i,j}\right) =$ $= 3\cdot 2^{k-2} - 1 + \left\lceil \log_2 \dfrac{k(k-1)}{2}\right\rceil$ (see [3, Section 5]).

On the other hand, there is the estimate based on the number of parallel straight lines containing the points of the support (see Proposition 4 in [3]). While the analytic complexity of any polynomial with the support belonging to a straight line does not exceed 2, the number of these lines corresponding to the i-th row of $\hat{A}$ equals $\hat{c}_i + 1$. Thus for any i the analytic complexity of the part of the solution whose support belongs to $\bigcup\limits_{j=1}^{k} \square_{i,j}$ cannot exceed $2 + \lceil \log_2(\max\limits_{i=1,\dots,k} \hat{c}_i + 1)\rceil$. Note that there is no need to use all of k pairs of bounding straight lines to estimate the analytic complexity of the general solution this way, since $k-1$ pairs already bound the whole support of the solution. The sum of $k-1$ elements in $Cl_{2+\lceil \log_2(\max\limits_{i=1,\dots,k} \hat{c}_i + 1)\rceil}$ cannot exceed $estim_2\left(\bigcup\limits_{i,j=1}^{k} \square_{i,j}\right) = 2 + \lceil \log_2\left(\max\limits_{i=1,\dots,k} \hat{c}_i + 1\right)\rceil + \lceil \log_2(k-1)\rceil$. The minimal of the numbers $estim_1\left(\bigcup\limits_{i,j=1}^{k} \square_{i,j}\right), estim_2\left(\bigcup\limits_{i,j=1}^{k} \square_{i,j}\right)$ is the sought estimate. □

An example of using the estimate given in Theorem 1 is shown in Fig. 3. Note that there are three sets of parallel lines, each corresponding to one of the $\hat{c}_i$. For each of the parallelogram intersections there are 2 estimates: $estim_1(\square_{i,j})$, based on Corollary 2 and $estim_2(\square_{i,j})$ based on Proposition 4 in [3].

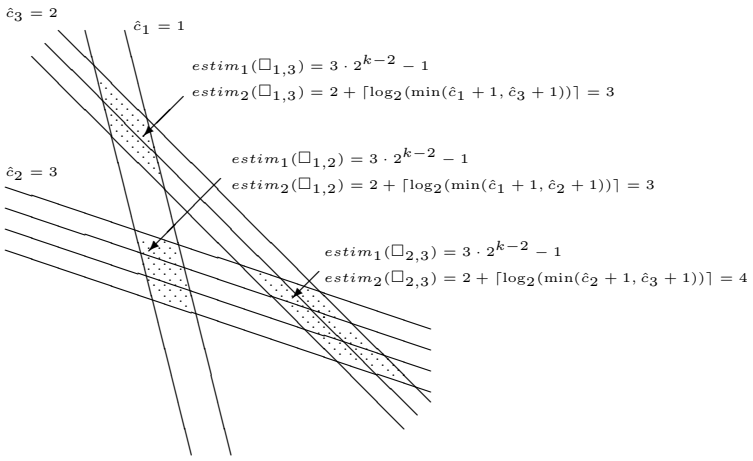


Fig. 3. The analytic complexity estimate for a polynomial solution to a simple hypergeometric system

We order $\hat{c}_i$ by the ascension and choose v to be a vector with the elements $v_i = \min\left(2 + \lceil \log_2(\hat{c}_i + 1) \rceil, 3 \cdot 2^{k-2} - 1 + \lceil \log_2(k - i) \rceil\right), i = 1, \ldots, k - 1$. To find more accurate value for the analytic complexity estimate from Theorem 1, one could use Algorithm 2 from Section 3 using v as an input vector. The general estimate from Theorem 1 can be rough, if values of $\hat{c}_i$ vary greatly for different i. For example in Fig. 3 $estim_2(\square_{1,2}) = estim_2(\square_{1,3}) = 3$, $estim_2(\square_{2,3}) = 4$, and for the general estimate we use the maximal of these values. The vector v in this case provides the choice of the better estimate.

3. Algorithms of analytic complexity estimation

The following algorithm allows one to compute the analytic complexity of any given bivariate polynomial.

Algorithm 1: Finding an analytic complexity estimate for a polynomial

Input: $p(x, y)$ - a polynomial, $x, y \in \mathbb{C}$.

Output: N - an estimate for the analytic complexity of $p(x, y)$.

```
result ← 0
short ← {}
polys ← {p_i(x,y) | p(x,y) = Σ_i p_i(x,y), Supp p_i(x,y) || Supp p_j(x,y) ∀i,j}
for p ∈ polys do
    curr = getShort(p)
    if curr ⊄ short then
        result += 1
    short = short ∪ curr
N ← 2 + ⌈Log_2(result)⌉
```

The main advantage of this algorithm compared to the existing ones is its ability to distinct the powers of lower degree polynomials included in the original polynomial as summands. Without this feature, even the analytic complexity of the function like $p(a(x) + b(y)) \in Cl_1$, where $p(t), a(x), b(y)$ are univariate polynomials, is estimated based on its support, which becomes very

inaccurate with the growth of degree of $p(t)$.

The input of the function *getShort*() is a homogeneous polynomial and the output contains elements of its decomposition into the sum of powers. Note that the definition of *polys* assumes the ambiguity of the representation of the polynomial as the sum of finitely many polynomials supported in parallel straight lines. Any of such representations yields an estimate, but some of them may be better than the other ones.

To estimate the analytic complexity of the general solution to the hypergeometric system from Theorem 1 one can use the following algorithm.

Algorithm 2: Finding an analytic complexity estimate for a sum of functions

Input: $c = \{c_1, c_2, \ldots, c_n\}$ - a set of known estimates of the analytic complexity values for bivariate functions $f_1(x,y), f_2(x,y), \ldots, f_n(x,y)$, where $(x,y) \in \mathbb{C}^2$.

Output: N - an estimate for the analytic complexity of the function $\sum_{i=1}^{n} f_i(x,y)$.

while c contains more than 1 element **do**
    find 2 minimal elements of c, namely, c_i and c_j.
    $c = (c \cup \{\max(c_i, c_j) + 1\}) \backslash \{c_i, c_j\}$.
$N \leftarrow$ only element of c.

Algorithm 2 is finite, since at each step the number of elements in c decreases by 1.

The following algorithm allows one to find the support of a polynomial solution to a given hypergeometric system defined by a zonotope, provided that such a solution exists. The algorithm is based on Proposition 4.7 in [10].

Algorithm 3: Constructing the support of a polynomial solution to a hypergeometric system

Input: the matrix A, the parameter vector c for the hypergeometric system $\mathrm{Horn}(A, c)$ defined by a zonotope

Output: *supp* - the support for the polynomial solution to $\mathrm{Horn}(A, c)$.

supp $\leftarrow \{\}$
find $\hat{A}$: $\mathrm{rows}(\hat{A}) \cup \mathrm{rows}(-\hat{A}) = \mathrm{rows}(A)$
for $(r_i, r_j) \subset \mathrm{rows}(\hat{A}), i < j$ **do**
    $A_{i,j} \leftarrow (r_i, r_j)^T$
    $\alpha \leftarrow$ elements of c corresponding to (r_i, r_j)
    $\beta \leftarrow$ elements of c corresponding to $(-r_i, -r_j)$
    if $-\alpha_j - \beta_j > 0$ *for* $j = 1, 2$ **then**
        $supp = supp \cup \mathrm{Supp}\left(x^{-A_{i,j}^{-1}\alpha}\left(1 + x^{-A_{i,j}^{-1}e_1}\right)^{-\alpha_1-\beta_1}\left(1 + x^{-A_{i,j}^{-1}e_2}\right)^{-\alpha_2-\beta_2}\right)$
    else
        the general solution to $\mathrm{Horn}(A, c)$ is not a polynomial

For some pairs of rows r_i, r_j the solution to the corresponding system defined by a parallelogram is not a polynomial. In this case, a part of the basis in the solution space can still consist of polynomials, and their supports can be found by means of Algorithm 3.

4. Examples

Example 3 *(Continued).* Let us replace the parameter vector c' in the system $\mathrm{Horn}(A', c')$ by

the vector $(k, 0, 0, 0, 0, 0)$. The corresponding system is given by

$$x\theta_x(\theta_x + \theta_y + k) - \theta_x(\theta_x + \theta_y),$$
$$y\theta_y(\theta_x + \theta_y + k) - \theta_y(\theta_x + \theta_y).$$

A basis in its solution space is given by 1, $\log \frac{x}{x-1} + \sum_{j=1}^{k-1} \frac{(-1)^j}{j(x-1)^j}$, $\log \frac{y}{y-1} + \sum_{j=1}^{k-1} \frac{(-1)^j}{j(y-1)^j}$, so there is no polynomial basis for these parameter values. Nevertheless, the analytic complexity of the general solution is equal to 1.

The present example shows that the analytic complexity of solutions to hypergeometric systems can be heavily dependent on parameter vectors defining these systems. A resonant choice of their parameters can drastically reduce the analytic complexity of general solutions to such systems.

Example 4. *An octagon zonotope.* Consider Example 6.8 in [10]. In order to find the analytic complexity of a polynomial solution to the hypergeometric system defined by the matrix

$$A = \begin{pmatrix} 1 & -1 & -1 & 1 & -3 & 3 & 2 & -2 \\ 2 & -2 & 1 & -1 & -2 & 2 & -1 & 1 \end{pmatrix}^{\mathrm{T}}$$

and the vector of parameters $c = (3, -5, -2, 1, -2, -1, -1, -1)$ we can use the basis of the solutions to this system, computed in [10]. There are 3 solutions whose analytic complexity is equal to 2, and 28 solutions in Cl_1, two of them also belonging to Cl_0. Therefore the analytic complexity of the general solution to this system cannot exceed 7. Note that this estimate is based on a trivial grouping of the basis functions into pairs, but the very specific structure of the solution support makes it possible to show that the analytic complexity does not exceed 6.

Let us estimate the analytic complexity of the general solution to this system using Theorem 1. The vector $\hat{c}$, ordered by the ascension, is $(1, 2, 2, 3)$. Then the vector $v = (3, 4, 4)$ (it includes only support-based estimates, because of low values of the elements of $\hat{c}$), and, by using Algorithm 2, we conclude that the general solution belongs to Cl_6.

Futhermore, we can estimate the analytic complexity of a solution to any hypergeometric system we obtain by appending a pair of rows to A (the only condition is that these rows are not collinear to the rows of A). Note that this estimate does not depend much on the difference between new parameters. If this difference is big, it becomes the last element of the ordered vector $\hat{c}$, and does not affect the new vector v, the new element of the vector v is equal to $2 + \lceil \log_2(3 + 1) \rceil = 4$, and the resulting analytic complexity is 6. On the contrary, if this difference is low, for example, if it is equal to 1, the new vector $\hat{c} = (1, 1, 2, 2, 3)$, the new vector $v = (3, 3, 4, 4)$, and the analytic complexity is also equal to 6. Thus we conclude that appending 2 rows to the matrix A does not affect the analytic complexity of the solution to the system.

Example 5. *A decagon zonotope.* Consider the hypergeometric system $\mathrm{Horn}(A_1, c_1)$, defined by the matrix

$$\begin{pmatrix} -1 & 1 & 0 & 0 & -2 & 2 & 3 & -3 & 3 & -3 \\ 0 & 0 & -1 & 1 & 1 & -1 & 1 & -1 & 2 & -2 \end{pmatrix}^{\mathrm{T}} \tag{3}$$

and the parameter vector $c_1 = (-1, 0, 4, -5, 1, -4, -9, 6, -4, 0)$. The zonotope defining the matrix 3 is shown in Fig. 4.

By Theorem 2.5 in [5] the holonomic rank of the system $\mathrm{Horn}(A_1, c_1)$ equals 34. The support to the solution to this system computed by the means of Algorithm 3 is shown in Fig. 5.

A polynomial basis in the solution space to $\mathrm{Horn}(A_1, c_1)$ consists of the 4 monomials $\frac{x^6}{y^9}, \frac{x^{17/3}}{y^8}, \frac{x^3}{y^3}, \frac{x^{8/3}}{y^2}$ and 30 polynomials

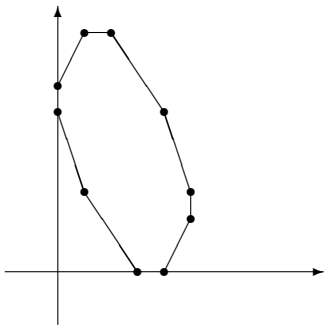


Fig. 4. The zonotope which defines the matrix (3)

$$\frac{1}{xy^6}+\frac{5643}{637xy^5}+\frac{247095}{8281xy^4}+\frac{329460}{8281xy^3}+\frac{27455}{286y^4}+\frac{82365}{49y^3}+\frac{741285}{49y^2}+\frac{724812}{7y},$$

$$\frac{20y^3}{63x}-\frac{4y^2}{35x}+\frac{4y^2}{5}-\frac{18y}{5}+1,\qquad \frac{3y^{3/2}}{380x}-\frac{3969y^{5/2}}{41990x}+\frac{1323y^{7/2}}{16796x}-\frac{51}{55}y^{3/2}+\sqrt{y},$$

$$-\frac{11y^{12}}{115311x}-\frac{33y^{11}}{38437x}-\frac{297y^{10}}{100555x}-\frac{24y^9}{5915x}+\frac{3y^9}{1105}+\frac{y^8}{26}+\frac{36y^7}{143}+y^6,\qquad xy^4-\frac{2}{13}xy^5,$$

$$\frac{8y^5}{99x}+\frac{4y^4}{3x}+\frac{50y^5}{81}+y^4,\qquad \frac{1550775x^{7/2}y^5}{82808479}-\frac{31465x^{9/2}y^5}{61400001}+x^{5/2}y^4-\frac{5175x^{7/2}y^4}{89947},$$

$$\frac{1547x^4y^5}{103455}-\frac{91x^4y^4}{6840}-\frac{91x^3y^5}{1026}+x^3y^4,\qquad \frac{806y^5}{129x^{8/3}}+\frac{84656y^4}{735x^{5/3}}+\frac{y^4}{x^{8/3}},\qquad \frac{x^{13/3}}{y^6}+\frac{451x^{13/3}}{261y^5},$$

$$\frac{87y^5}{82x^{7/3}}+\frac{5220y^5}{275561x^{10/3}}+\frac{36575y^4}{2392x^{4/3}}+\frac{y^4}{x^{7/3}},\qquad \frac{44y^5}{1183x^3}+\frac{33y^5}{182x^2}+\frac{y^4}{x^2},\qquad \frac{x^{16/3}}{y^8}+\frac{1378x^{16/3}}{451y^7},$$

$$-\frac{21}{46}x^{2/3}y^5+x^{2/3}y^4+\frac{119}{286}x^{5/3}y^4,\qquad -\frac{12}{247}x^{4/3}y^5+x^{4/3}y^4-\frac{364\sqrt[3]{x}y^5}{1045},\qquad \frac{2x}{7y}+x,$$

$$\frac{11985}{299}x^{8/7}y^{2/7}+\frac{14382x^{8/7}}{253y^{5/7}}+\frac{x^{8/7}}{y^{12/7}},\qquad \frac{1200x^{2/7}}{1643y^{3/7}}+\frac{345x^{9/7}}{31y^{3/7}}+\frac{x^{9/7}}{y^{10/7}},\qquad \frac{x^{10/3}}{y^4}+\frac{261x^{10/3}}{238y^3},$$

$$\frac{114774x^{6/7}y^{5/7}}{28405}+\frac{1188x^{6/7}}{65y^{2/7}}+\frac{x^{6/7}}{y^{9/7}},\qquad \frac{x^{10/7}}{y^{8/7}}+\frac{731x^{3/7}}{638\sqrt[7]{y}}+\frac{1763x^{10/7}}{754\sqrt[7]{y}},$$

$$\frac{x^{4/7}}{y^{6/7}}+\frac{32680x^{11/7}}{8613y^{6/7}}+\frac{1558}{261}x^{4/7}\sqrt[7]{y},\qquad \frac{169}{150}x^{5/7}y^{3/7}+\frac{x^{5/7}}{y^{4/7}}+\frac{65x^{12/7}}{136y^{4/7}},$$

$$-\frac{1}{66}5x^2y^3+\frac{5}{7}x^2y^2-\frac{45}{28}x^2y+x^2,\qquad x^{11/5}y^{2/5}-\frac{4301x^{11/5}y^{7/5}}{4277}+\frac{232254x^{11/5}y^{12/5}}{1056419},$$

$$x^{9/5}y^{3/5}-\frac{1287x^{9/5}y^{8/5}}{1634}+\frac{55913x^{9/5}y^{13/5}}{346408},\qquad x^{12/5}y^{4/5}-\frac{68}{19}x^{7/5}y^{9/5}-\frac{116}{231}x^{12/5}y^{9/5},$$

$$x^{8/5}y^{6/5}+\frac{5824x^{13/5}y^{6/5}}{432837}-\frac{1064x^{8/5}y^{11/5}}{2829},\qquad \frac{x^5}{y^7}+\frac{8x^5}{15y^6}-\frac{21x^4}{55y^6}-\frac{182x^4}{15y^5}-\frac{91x^4}{24y^4},$$

$$\frac{x^{14/3}}{y^7}+\frac{828x^{14/3}}{85y^6}-\frac{585488x^{11/3}}{48825y^5}+\frac{21758x^{14/3}}{23715y^5}-\frac{2488324x^{11/3}}{35805y^4}.$$

There are 14 functions in Cl_1 and 20 functions in $Cl_2\backslash Cl_1$ among these polynomials.

The analytic complexity estimate of the general solution to $\mathrm{Horn}(A_1, c_1)$ obtained by grouping these functions into pairs is Cl_7. Theorem 1 gives a better estimate: since $\hat{c} = (2, 2, 3, 3, 4)$,

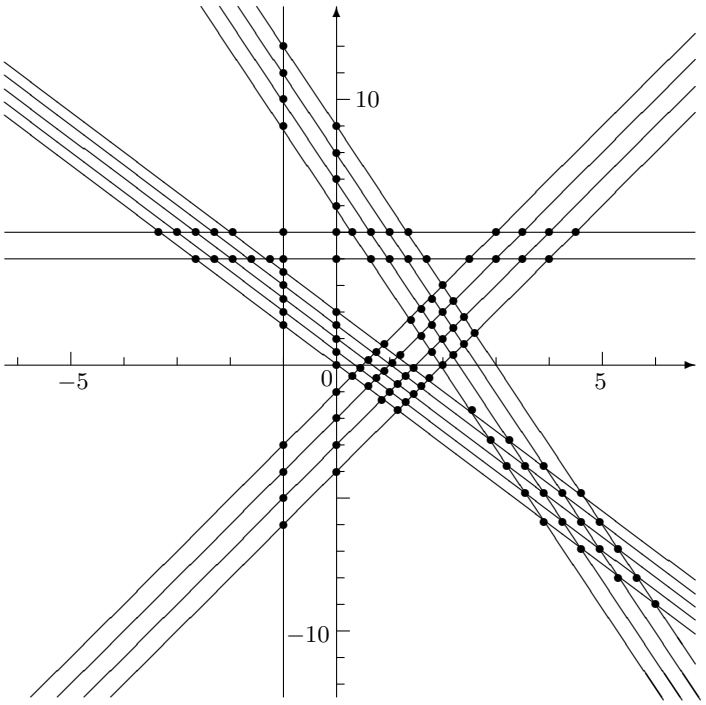


Fig. 5. The support for the solution of the system $\mathrm{Horn}(A_1, c_1)$

$v = (4, 4, 4, 4)$, it follows that the general solution belongs to Cl_6.

The following examples present hypergeometric systems defined by polygons other than zonotopes whose solutions have low analytic complexity.

Example 6. *A pentagon.* The matrix $\begin{pmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & -1 & 1 \end{pmatrix}^{\mathrm{T}}$ and the vector of parameters $(-4, 0, 0, -1, -2, -1, -2)$ define the hypergeometric system

$$\begin{aligned} &x(\theta_x + \theta_y - 4)(\theta_x - 1) - \theta_x(\theta_x - 2), \\ &y(\theta_x + \theta_y - 4)(\theta_y - 1) - \theta_y(\theta_y - 2). \end{aligned} \tag{4}$$

This system is holonomic and its holonomic rank equals 4. The pure basis (see [10]) in its solution space is given by the Taylor polynomials

$$x^2y^2, \quad 1 - 4x - 4y + 12xy, \quad 6x^2 - 4x^3 + x^4 - 12x^2y + 4x^3y, \quad 6y^2 - 12xy^2 - 4y^3 + 4xy^3 + y^4.$$

The first and the second of these polynomials belong to Cl_1, the third and the fourth belong to Cl_2. Thus the general solution is a function in Cl_4.

Example 7. *A trapezoid, high holonomic rank.* A basis in the solution space of the hypergeometric system with holonomic rank k defined by the operators

$$\begin{aligned} &x\theta_x^{k-1}(\theta_x + \theta_y) - (-1)^k\theta_x^k, \\ &y(\theta_x + \theta_y) + \theta_y. \end{aligned}$$

is given by $\{\log^j((y+1)/x),\ j = 0, \ldots, k-1\}$ (see Fig. 6). The generating solution equals $\log^{k-1}((y+1)/x)$. Thus the general solution to this system belongs to Cl_1 by the conservation principle. This example shows that the analytic complexity of solutions to hypergeometric systems with high holonomic rank can still be low.

Fig. 6. a) the supports of solutions to the system (4); b) polygon defining the system (4)

Example 8. *A triangle with no symmetries.* The hypergeometric system

$$\begin{aligned} &x(\theta_x+\theta_y-4)(\theta_x+2\theta_y-4)-(2\theta_x+3\theta_y-4)(2\theta_x+3\theta_y-5),\\ &y(\theta_x+\theta_y-4)(\theta_x+2\theta_y-4)(\theta_x+2\theta_y-3)-(2\theta_x+3\theta_y-4)(2\theta_x+3\theta_y-5)(2\theta_x+3\theta_y-6) \end{aligned} \tag{5}$$

is holonomic and its holonomic rank equals 6. The pure basis in its solution space is given by the Laurent polynomials

$$x^{-4}y^4,\quad x^{-2}y^3,\quad x^7y^{-3},\quad x^8y^{-4},\quad 3y^2+2x^{-1}y^2,$$

$$6x^2+12x^3+x^4+4x^5y^{-2}+6x^6y^{-2}-12x^4y^{-1}-4x^5y^{-1}-12xy-4x^2y.$$

In the Fig. 7 the small filled circles correspond to monomial solutions, the two empty circles indicate the binomial solution and the big filled circles correspond to the remaining polynomial solution. The analytic complexity of the general solution to the system (7) does not exceed 5.

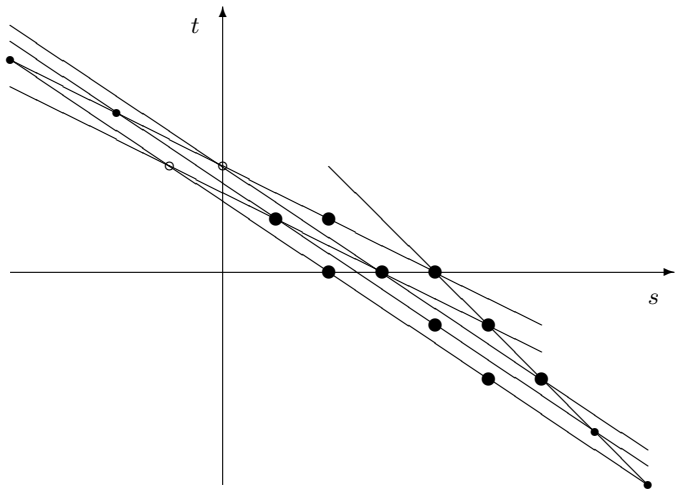


Fig. 7. The supports of solutions to the system (5)

This research was performed in the framework of the state task in the field of scientific activity of the Ministry of Science and Higher Education of the Russian Federation, project "Development of the methodology and a software platform for the construction of digital twins, intellectual analysis and forecast of complex economic systems", Grant no. FSSW-2020-0008.

References

[1] V.I.Arnold, On the representation of continuous functions of three variables by superpositions of continuous functions of two variables, *Sbornik Mathematics*, **48**(1959), no. 1, 3–74.

[2] V.K.Beloshapka, Analytic complexity of functions of two variables, *Russian J. Math. Phys.*, **14**(2007), no. 3, 243–249.

[3] V.K.Beloshapka, Analytical complexity: Development of the topic, *Russian J. Math. Phys.*, **19**(2012), no.4, 428–439.

[4] V.K.Beloshapka, On the complexity of differential algebraic definition for classes of analytic complexity, *Math. Notes*, **105**(2019), no. 3, 323–331.

[5] A.Dickenstein, L.F.Matusevich, T.M.Sadykov, Bivariate hypergeometric D-Modules, *Advances in Mathematics*, **196**(2005), 78–123.

[6] A.Dickenstein, T.M.Sadykov, Algebraicity of solutions to the Mellin system and its monodromy, *Dokl. Math.*, **75**(2007), no. 1, 80–82. DOI: 10.1134/S106456240701022X

[7] A.Dickenstein, T.M.Sadykov, Bases in the solution space of the Mellin system, *Sbornik Mathematics*, **198**(2007), no. 9, 1277–1298.

[8] J.Horn, Über die Konvergenz der hypergeometrischen Reihen zweier und dreier Veränderlichen, *Math. Ann.*, **34**(1889), 544–600.

[9] V.A.Krasikov, Analytic complexity of hypergeometric functions satisfying systems with holonomic rank two, *Lecture Notes in Computer Science*, **11661**(2019), 330–342.

[10] T.M.Sadykov, S.Tanabe, Maximally reducible monodromy of bivariate hypergeometric systems, *Izv. Math.*, **80**(2016), no. 1, 221–262.

[11] T.M.Sadykov, Beyond the first class of analytic complexity, *Lecture Notes in Computer Science*, **11077**(2018), 335–344.

[12] T.M.Sadykov, Computational problems of multivariate hypergeometric theory, *Programming and Computer Software*, **44**(2018), no. 2, 131–137. DOI: 10.1134/S0361768818020093

[13] T.M.Sadykov, The Hadamard product of hypergeometric series, *Bulletin des Sciences Mathematiques*, **126**(2002), no. 1, 31.

[14] T.M.Sadykov, On the analytic complexity of hypergeometric functions, *Proceedings of the Steklov Institute of Mathematics*, **298**(2017), no. 1, 248–255.
DOI: 10.1134/S0081543817060165

[15] M.A.Stepanova, Analytic complexity of differential algebraic functions, *Sbornik Mathematics*, **210**(2019), no. 12, 1774–1787.

[16] M.A.Stepanova, On analytical complexity of antiderivatives, *Journal of Siberian Federal University. Mathematics & Physics*, **12**(2019), no. 6, 694–698.
DOI: 10.17516/1997-1397-2019-12-6-694-698

[17] A.G.Vitushkin, On Hilbert's thirteenth problem and related questions, *Russian Math. Surveys*, **59**(2004), no. 1, 11–25.

Верхние границы аналитической сложности решений двумерных гипергеометрических систем в классе многочленов Пюизо

Виталий А. Красиков
Российский экономический университет им. Г. В. Плеханова
Москва, Российская Федерация

Аннотация. В статье исследуется аналитическая сложность решений двумерных голономных гипергеометрических систем типа Горна. Получены оценки аналитической сложности решений в классе многочленов Пюизо для гипергеометрических систем, заданных зонотопами. Также предложены алгоритмы для оценки аналитической сложности многочленов.

Ключевые слова: гипергеометрические системы дифференциальных уравнений в частных производных, голономный ранг, полиномиальные решения, зонотопы, аналитическая сложность, дифференциальный многочлен.

DOI: 10.17516/1997-1397-2020-13-6-733-745
УДК 512.554.38

Almost Inner Derivations of Some Nilpotent Leibniz Algebras

Zhobir K. Adashev*
Institute of Mathematics of Uzbek Academy of Sciences
Tashkent, Uzbekistan
Tuuelbay K. Kurbanbaev†
Karakalpak State University
Nukus, Uzbekistan

Received 06.07.2020, received in revised form 08.08.2020, accepted 16.10.2020

Abstract. We investigate almost inner derivations of some finite-dimensional nilpotent Leibniz algebras. We show the existence of almost inner derivations of Leibniz filiform non-Lie algebras differing from inner derivations, we also show that the almost inner derivations of some filiform Leibniz algebras containing filiform Lie algebras do not coincide with inner derivations.

Keywords: Leibniz algebra, derivation, inner derivation, almost inner derivation.

Citation: J.K. Adashev, T.K. Kurbanbaev, Almost Inner Derivations of some Nilpotent Leibniz Algebras, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 733–745. DOI: 10.17516/1997-1397-2020-13-6-733–745.

Introduction

Lie algebra is an algebra satisfying the anticommutativity identity and the Jacobi identity. The derivations of finite-dimensional Lie algebras are a well-studied direction of the theory of Lie algebras. It should be noted that the space of all derivations of Lie algebras is also Lie algebra with respect to the commutator. In the set of derivations of Lie algebras, there exist subsets of the so-called inner derivations. Naturally, there is a question: in what classes of algebras do derivations exist? and which are not inner? For the semisimple Lie algebras the sets of inner derivations and derivations coincide [14].

Almost inner derivations of Lie algebras were introduced by C. S. Gordon and E. N. Wilson [13] in the study of isospectral deformations of compact manifolds. Gordon and Wilson wanted to construct not only finite families of isospectral nonisometric manifolds, but rather continuous families. They constructed isospectral but nonisometric compact Riemannian manifolds of the form G/Γ, with a simply connected exponential solvable Lie group G, and a discrete cocompact subgroup Γ of G. For this construction, almost inner automorphisms and almost inner derivations were crucial.

Gordon and Wilson considered not only almost-inner derivations, but they studied almost inner automorphisms of Lie groups. The concepts of "almost inner" automorphisms and derivations, almost homomorphisms or almost conjugate subgroups arise in many contexts in algebra, number theory and geometry. There are several other studies of related concepts, for example, local derivations, which are a generalization of almost inner derivations and automorphisms [2,3].

In [4] we initiated the study of derivation type maps on non-associative algebras, namely, we investigated so-called 2-local derivations on finite-dimensional Lie algebras, and showed an essential difference between semisimple and nilpotent Lie algebras is the behavior of their 2-local

*adashevjq@mail.ru https://orcid.org/0000-0002-4882-4098
†tuuelbay@mail.ru https://orcid.org/0000-0002-9963-872X

derivations. The present paper is devoted to local derivation on finite-dimensional Lie algebra over an algebraically closed field of characteristic zero.

Local derivation first was considered in 1990, Kadison [16] and Larson and Sourour [18]. Let X be a Banach A-bimodule over a Banach algebra A, a linear mapping $\Delta : A \to X$ is said to be a *local derivation* if for every x in A there exists a derivation $D_x : A \to X$, depending on x, satisfying $\Delta(x) = D_x(x)$.

The main problems concerning this notion are to find conditions under which local derivations become derivations and to present examples of algebras with local derivations that are not derivations [8, 16, 18]. Kadison proves in [16, Theorem A] that each continuous local derivation of a von Neumann algebra M into a dual Banach M-bimodule is a derivation. This theorem gave rise to studies and several results on local derivations on C*-algebras, culminating with a definitive contribution due to Johnson, which asserts that every continuous local derivation of a C*-algebra A into a Banach A-bimodule is a derivation [15, Theorem 5.3]. Moreover in his paper, Johnson also gives an automatic continuity result by proving that local derivations of a C*-algebra A into a Banach A-bimodule X are continuous even if not assumed a priori to be so (cf. [15, Theorem 7.5]).

In the theory of Lie algebras, there is a theorem which says that in the finite-dimensional nilpotent Lie algebra there are not inner (i.e. outer) derivations [12]. We give an Example 2.1 to shows that that there exists 4-dimensional nilpotent Lie algebras, where any almost inner derivation is an outer derivation, and the converse is true also. But this question is still open for the general case. In [9] authors study almost inner derivations of some nilpotent Lie algebras. Prove the basic properties of almost inner derivations, calculate all almost inner derivations of Lie algebras for small dimensions. They also introduced the concept of fixed basis vectors for nilpotent Lie algebras defined by graphs and studied free nilpotent Lie algebras of the nilindex 2 and 3.

We recall that the study of almost-inner derivations of the Leibniz algebras is an open problem. Therefore in this paper we consider almost-inner derivations for some nilpotent Leibniz algebras. We prove the basic properties of almost inner derivations of the Leibniz algebras. We get almost all inner derivations of four-dimensional nilpotent Leibniz algebras. The study of the inner derivations of nilpotent Leibniz algebras is a very difficult problem. Therefore, we consider some subclasses of these nilpotent algebras. We study almost inner derivations of the null-filiform Leibniz algebras, and also consider almost inner derivations of the some filiform Leibniz algebras.

1. Preliminaries

Definition 1.1. *An algebra L over a field F is called the* Leibniz algebra *if for all $x, y, z \in L$ the Leibniz identity holds:*

$$[x,[y,z]] = [[x,y],z] - [[x,z],y],$$

where $[\ ,\ ]$ is the multiplication in L.

For an arbitrary Leibniz algebra L, we define a sequence:

$$L^1 = L, \quad L^{k+1} = [L^k, L^1], \quad k \geqslant 1.$$

The Leibniz algebra L is said to be *nilpotent* if there exists $s \in \mathbb{N}$ such that $L^s = 0$. The minimal number s with this property is called the *nilpotency index* or *nilindex* of the algebra L.

We recall that the Leibniz algebra is called

null-filiform, if $dim L^i = (n+1) - i,\ 1 \leqslant i \leqslant n+1$;

filiform, if $dim L^i = n - i,\ 2 \leqslant i \leqslant n$.

Let L be a nilpotent Leibniz algebra with nilindex s.

We consider $L_i = L^i/L^{i+1}$, $1 \leqslant i \leqslant s-1$ and $grL = L_1 \oplus L_2 \oplus \cdots \oplus L_{s-1}$. Then $[L_i, L_j] \subseteq L_{i+j}$ and we obtain the graded algebra grL.

Definition 1.2. *If the Leibniz algebra L is isomorphic algebra grL, then L is called naturally graded Leibniz algebra.*

For the Leibniz algebra L, we denote the *right* and *left* annihilators, respectively, as follows

$$Ann_r(L) = \{x \in L \mid [L, x] = 0\}, \quad Ann_l(L) = \{x \in L \mid [x, L] = 0\}.$$

We denote the *center* of the algebra by $Cent(L) = Ann_r(L) \cap Ann_l(L)$.

A linear map d is called a *derivation* of the Leibniz algebra L, if

$$d\big([x, y]\big) = [d(x), y] + [x, d(y)].$$

We denote the space of all derivations by $Der(L)$.

For each $x \in L$, the operator $R_x : L \to L$ which is called the *right multiplication*, such that $R_x(y) = [y, x]$, $y \in L$, is a derivation. This derivation is called *an inner derivation* of L, and we denote the space of all inner derivations by $Inner(L)$.

Definition 1.3. *The derivation $D \in Der(L)$ of the Leibniz algebra L is called almost inner derivation, if $D(x) \in [x, L]$ $\big([x, L] \subseteq L\big)$ holds for all $x \in L$; in other words, there exists $a_x \in L$ such that $D(x) = [x, a_x]$.*

The space of all almost inner derivations of L is denoted by $AID(L)$.

Definition 1.4. *The derivation $D \in AID(L)$ of the Leibniz algebra L is called **the right central almost inner derivation**, if there exists $x \in L$ such that the map $(D - R_x) : L \to Ann_r(L)$.*

The space of right central almost inner derivations of L is denoted by $RCAID(L)$, respectively.

Definition 1.5. *The derivation $D \in AID(L)$ of the Leibniz algebra L is called **central almost inner derivation**, if there exists $x \in L$ such that the map $(D - R_x) : L \to Cent(L)$.*

The space of central almost inner derivations of L is denoted by $CAID(L)$, respectively.

2. Main results

2.1. The properties of almost inner derivations of the Leibniz algebras

The subspaces $Inner(L)$, $CAID(L)$, $RCAID(L)$, $AID(L)$, $Der(L)$ are Lie subalgebras with $[D, D'] = DD' - D'D$.

Proposition 2.1. *We have the following inclusions of Lie subalgebras*

$$Inner(L) \subseteq CAID(L) \subseteq RCAID(L) \subseteq AID(L) \subseteq Der(L).$$

Proof. Let $D_1, D_2 \in AID(L)$ and $x \in L$. Then there exist $y_1, y_2 \in L$ such that $D_1(x) = [x, y_1]$, $D_2(x) = [x, y_2]$. Using the property of the derivation and the Leibniz identity, we get the following

$$\begin{aligned}[D_1, D_2](x) = (D_1D_2)(x) - (D_2D_1)(x) &= [D_1(x), y_2] + [x, D_1(y_2)] - [D_2(x), y_1] - [x, D_2(y)] = \\ &= [[x, y_1], y_2] - [[x, y_2], y_1] + [x, D_1(y_2)] - [x, D_2(y_1)] = \\ &= [x, [y_1, y_2]] + [x, D_1(y_2)] - [x, D_2(y_1)] = [x, [y_1, y_2] + D_1(y_2) - D_2(y_1)].\end{aligned}$$

Therefore, $[D_1, D_2](x) = [x, [y_1, y_2] + D_1(y_2) - D_2(y_1)] \in [x, L]$, we have $[D_1, D_2] \in AID(L)$.

Let $C_1, C_2 \in CAID(L)$. Then there exist $y_1, y_2 \in L$ such that $C_1 - R_{y_1}$ and $C_2 - R_{y_2}$ are maps from L to $Cent(L)$. We consider $[C, R_x] = R_{C(x)}$ for $C \in Der(L)$ and obtain the following

$$[C_1 - R_{y_1}, C_2 - R_{y_2}] = [C_1, C_2] - [C_1, R_{y_2}] - [R_{y_1}, C_2] + [R_{y_1}, R_{y_2}] =$$
$$= [C_1, C_2] - R_{C_1(y_2)} + R_{C_2(y_1)} - R_{[y_2,y_1]} = [C_1, C_2] - (R_{C_1(y_2)} - R_{C_2(y_1)} + R_{[y_2,y_1]}).$$

Hence we have that the linear transformation $[C_1, C_2] - (R_{C_1(y_2)} - R_{C_2(y_1)} + R_{[y_2,y_1]})$ maps L to $Cent(L)$. Hence $[C_1, C_2] \in CAID(L)$.

Let $D_1, D_2 \in RCAID(L)$. Then there exist $y_1, y_2 \in L$ such that $D_1 - R_{y_1}$ and $D_2 - R_{y_2}$ are maps L to $Ann_r(L)$. We consider $[D, R_x] = R_{D(x)}$ for $D \in Der(L)$ and obtain the following

$$[D_1 - R_{y_1}, D_2 - R_{y_2}] = [D_1, D_2] - [D_1, R_{y_2}] - [R_{y_1}, D_2] + [R_{y_1}, R_{y_2}] =$$
$$= [D_1, D_2] - R_{D_1(y_2)} + R_{D_2(y_1)} - R_{[y_2,y_1]} = [D_1, D_2] - (R_{D_1(y_2)} - R_{D_2(y_1)} + R_{[y_2,y_1]}).$$

Hence we have that the linear transformation $[D_1, D_2] - (R_{D_1(y_2)} - R_{D_2(y_1)} + R_{[y_2,y_1]})$ maps L to $Ann_r(L)$. Hence $[D_1, D_2] \in RCAID(L)$.

Now let us show that $Inner(L) \subseteq CAID(L)$. Let $R_x, R_y \in Inner(L)$ and $R_x - R_y : L \to Cent(L)$. For every $z \in L,\ a \in Cent(L)$ we consider the following

$$(R_x - R_y)(z) = [z, x] - [z, y] = [z, x] - [z, a + x] = [z, a] \in Cent(L).$$

Therefore, $Inner(L) \subseteq CAID(L)$. □

Proposition 2.2. *The subalgebra $RCAID(L)$ is a Lie ideal in $AID(L)$.*

Proof. Let $C \in RCAID(L)$ and $D \in AID(L)$. We must show $[D, C] \in RCAID(L)$. We already know $[D, C] \in AID(L)$. We fix an element $x \in L$ such that $C' := C - R_x$ maps L to $Ann_r(L)$. We denote $D' := [D, C] - R_{D(x)}$. Then from $[D, R_x] = R_{D(x)}$ we obtain

$$[D, C'] = [D, C - R_x] = [D, C] - [D, R_x] = [D, C] - R_{D(x)} = D'$$

and D' maps L to $Ann_r(L)$. Hence for all $y \in L$ we have

$$D'(y) = [D, C'](y) = D(C'(y)) - C'(D(y)),$$

because C' maps L to $Ann_r(L)$ and D maps $Ann_r(L)$ to $Ann_r(L)$. □

Proposition 2.3. *Let L be the Leibniz algebra. Then the followings are true:*

1) Let $D \in AID(L)$. Then $D(L) \subseteq [L, L]$, $D(Cent(L)) = 0$ and $D(I) \subseteq I$ for every ideal I of L.

2) For $D \in CAID(L)$, there exists an $x \in L$ such that $D_{|[L,L]} = R_{x|[L,L]}$.

3) If L has nilindex 3, then $CAID(L) = AID(L)$.

4) If $Cent(L) = 0$, then $CAID(L) = Inner(D)$.

5) If L is nilpotent, then $AID(L)$ is nilpotent.

6) $AID(L \oplus L') = AID(L) \oplus AID(L')$.

Proof. 1) By definition, almost inner derivations of L maps to $[L, L]$ and $Cent(L)$ to 0.

Let $x \in I$. Then we have $D(x) \in [x, L] \subseteq [I, L] \subseteq I$.

2) For a given $D \in CAID(L)$, there exists $x \in L$ such that $D' = D - R_x$ satisfies $D'(L) \subseteq Cent(L)$. Hence D' is derivation and for all $u, v \in L$ we have

$$D'\big([u, v]\big) = [D'(u), v] + [u, D'(v)] = 0.$$

3) If L is nilpotent with nilindex 3, i.e. $L^3 = 0$, then for each $D \in AID(L)$ we get $D(L) \subseteq [L, L] \subseteq Cent(L)$ and get equality.

4) We suppose $Cent(L) = 0$ and $D \in CAID(L)$. Then there is $x \in L$ such that $D - R_x = 0$. Therefore D is inner.

5) Let $D \in AID(L)$ and $x \in L$. Then $D^k(x) \in [[[\ldots, [x, L], \ldots L], L]$ (k times L). If k is higher than nilpotent class over L, then we have $D^k(x) = 0$, therefore D is nilpotent. By Engel's theorem for Leibniz algebras [5], $AID(L)$ is nilpotent.

6) Let $D \in AID(L \oplus L')$. Then the constraints are again almost inner derivations, i.e. $D_{|L} \in AID(L)$ and $D_{|L'} \in AID(L')$. It is obvious that the mapping $D \mapsto D_{|L} \oplus D_{|L'}$ gives a one-to-one correspondence between $AID(L \oplus L')$ and $AID(L) \oplus AID(L')$.

□

2.2. Almost inner derivations of null-filiform Leibniz algebras

Firstly we consider a certain class of nilpotent Leibniz algebras, the so-called null-filiform Leibniz algebra [7].

In any n-dimensional null-filiform Leibniz algebra L there exists a basis $\{e_1, e_2, \ldots, e_n\}$ such that the multiplication in L has the form:

$$NF_n: \quad [e_i, e_1] = e_{i+1}, \quad 1 \leqslant i \leqslant n-1 \tag{1}$$

(the omitted of products are equal to zero).

Let L be a null-filiform Leibniz algebra.

Proposition 2.4. *For the n-dimensional null-filiform Leibniz algebra NF_n the following equality holds:*

$$AID(NF_n) = Inner(NF_n).$$

Proof. The null-filiform algebra L is a one-generated algebra, i.e. generated by e_1. Let $D \in AID(NF_n)$. Then, by the definition of almost inner derivation, there exists a_{e_1} such that $D(e_1) = R_{a_{e_1}}$. Let $D' \in AID(NF_n)$ and let $D' = D - R_{a_{e_1}}$, then we get $D'(e_1) = 0$. Then by multiplication (1) we have

$$D'(e_i) = D'\big([e_{i-1}, e_1]\big) = [D'(e_{i-1}, e_1)] + [e_{i-1}, D'(e_1)] = 0, \quad 2 \leqslant i \leqslant n.$$

This means that

$$AID(NF_n) = Inner(NF_n).$$

□

2.3. Almost inner derivation of non-lie filiform Leibniz algebras

Now we consider filiform non-Lie Leibniz algebras $F_1(\alpha_4, \alpha_5, \ldots, \alpha_n, \theta)$ and $F_2(\beta_5, \ldots, \beta_n, \gamma)$ from [7]:

$$F_1(\alpha_4, \alpha_5, \ldots, \alpha_n, \theta): \begin{cases} [e_1, e_1] = e_3, \\ [e_i, e_1] = e_{i+1}, \quad 2 \leqslant i \leqslant n-1, \\ [e_1, e_2] = \sum\limits_{s=4}^{n-1} \alpha_s e_s + \theta e_n, \\ [e_j, e_2] = \sum\limits_{s=4}^{n-j+2} \alpha_s e_{s+j-2}, \quad 2 \leqslant j \leqslant n-2, \end{cases}$$

$$F_2(\beta_4,\beta_5,\dots,\beta_n,\gamma):\begin{cases}[e_1,e_1]=e_3,\\ [e_i,e_1]=e_{i+1},\quad 3\leqslant i\leqslant n-1,\\ [e_1,e_2]=\sum\limits_{k=4}^{n}\beta_k e_k,\\ [e_2,e_2]=\gamma e_n,\\ [e_i,e_2]=\sum\limits_{k=4}^{n+2-i}\beta_k e_{k+i-2},\quad 3\leqslant i\leqslant n-2.\end{cases}$$

Let L be an algebra from $F_1(\alpha_4,\alpha_5,\dots,\alpha_n,\theta)$ or $F_2(\beta_4,\beta_5,\dots,\beta_n,\gamma)$.

Let L be the Leibniz algebra and $E_{n,2}:L\to L$ be a linear mapping such that

$$E_{n,2}(e_i)=\delta_{i,2}e_n,\quad 1\leqslant i\leqslant n, \tag{2}$$

where $\delta_{i,2}=\begin{cases}1, & i=2\\ 0, & i\neq 2\end{cases}$ — Kronecker symbol.

Theorem 2.1. *Let L be a non-Lie filiform Leibniz algebra and let $D\in AID(L)$. Then there exist an element $x\in L$ and $\lambda\in\mathbb{C}$ such that*

$$D-R_x=\lambda E_{n,2}.$$

Proof. We first consider the non-Lie filiform Leibniz algebra $L=F_1(\alpha_4,\alpha_5,\dots,\alpha_n,\theta)$.

Let $D\in AID(L)$. This algebra is a two-generated algebra, i.e. we have generators e_1 and e_2. Then, by the definition of almost inner derivation, there exists a_{e_1} such that $D(e_1)=R_{a_{e_1}}$. Let $D'\in AID(L)$ and $D'=D-R_{a_{e_1}}$, then we get $D'(e_1)=0$. Since $D'(e_1)=0$, then we have the following:

$$\begin{aligned}&D'(e_3)=D'\big([e_1,e_1]\big)=[D'(e_1),e_1]+[e_1,D'(e_1)]=0,\\ &D'(e_i)=D'\big([e_{i-1},e_1]\big)=[D'(e_{i-1}),e_1]+[e_{i-1},D'(e_1)]=[D'(e_{i-1}),e_1]=0,\quad 4\leqslant i\leqslant n.\end{aligned}$$

Let $D'(e_2)=\sum\limits_{j=1}^{n}b_je_j$. we check the following:

$$D'(e_3)=D'\big([e_2,e_1]\big)=[D'(e_2),e_1]=\left[\sum_{j=1}^{n}b_je_j,e_1\right]=(b_1+b_2)e_3+b_3e_4+\cdots+b_{n-1}e_n.$$

On the other hand, $D'(e_3)=D([e_1,e_1])=0$. So we get

$$b_1=-b_2,\quad b_i=0,\ 3\leqslant i\leqslant n-1.$$

Now we check the following:

$$\begin{aligned}0&=D'\big([e_1,e_2]\big)=[D'(e_1),e_2]+[e_1,D'(e_2)]=[e_1,b_1e_1-b_1e_2+b_ne_n]=\\ &=b_1e_3-b_1(\alpha_4e_4+\cdots+\alpha_{n-1}e_{n-1}+\theta e_n).\end{aligned}$$

We have $b_1=0$ and $D'(e_2)=b_ne_n$. On the other hand, by definition of almost inner derivation

$$b_ne_n=D'(e_2)=[e_2,a_{e_2}]=[e_2,a_{2,1}e_1+a_{2,2}e_2+\cdots+a_{2,n}e_n]=a_{2,1}e_3+a_{2,2}(\alpha_4e_4+\alpha_5e_5+\cdots+\alpha_ne_n).$$

We obtain

$$\begin{cases}a_{2,1}=0,\quad a_{2,2}\alpha_i=0,\ 4\leqslant i\leqslant n-1,\\ b_n=a_{2,2}\alpha_n.\end{cases} \tag{3}$$

Hence $D'(e_2) = a_{2,2}\alpha_n e_n$. If $a_{2,2}\alpha_n = 0$, then $AID(L) = Inner(L)$, so

$$a_{2,2}\alpha_n \neq 0,$$

therefore from (3) we get

$$\alpha_i = 0, \quad 4 \leqslant i \leqslant n-1.$$

In the end we obtain $D' = a_{2,2}\alpha_n E_{n,2} = \lambda E_{n,2}$.

Let $L = F_2(\beta_4, \beta_5, \ldots, \beta_n, \gamma)$ and $D' \in AID(L)$. By definition AID for e_2 there exists a_{e_2} such that

$$D'(e_2) = [e_2, a_{e_2}] = [e_2, a_{2,1}e_1 + \cdots + a_{2,n}] = a_{2,2}\gamma e_n.$$

Conducting analogously reasoning in this algebra we obtain $D'(e_1) = 0$, $D'(e_i) = 0$, $3 \leqslant i \leqslant n$ and $D' = a_{2,2}\gamma E_{n,2} = \lambda E_{n,2}$, where $\lambda \in \mathbb{C}$.

Now we consider the following equality:

$$\begin{aligned} a_{2,2}\gamma e_n &= D'(e_1) + D'(e_2) = D'(e_1 + e_2) = [e_1 + e_2, c_{e_1+e_2}] = [e_1 + e_2, c_1e_1 + c_2e_2] = \\ &= c_1e_3 + c_2\beta_4 e_4 + c_2(\beta_4 + \beta_5)e_6 + \cdots + c_2(\beta_4 + \cdots + \beta_{n-1})e_{n-1} + \\ &\quad + c_2(\beta_4 + \cdots + \beta_{n-1} + \beta_n + \gamma)e_n. \end{aligned}$$

We get

$$\begin{cases} c_1 = 0, \\ c_2\beta_i = 0, \ 4 \leqslant i \leqslant n-1, \\ c_2(\beta_n + \gamma) = a_{2,2}\gamma. \end{cases}$$

If at least one of $\beta_{i_0} \neq 0$ $(4 \leqslant i_0 \leqslant n-1)$, then we have $c_2 = 0$, hence $AID(L) = Inner(L)$. Therefore $\beta_i = 0$, $4 \leqslant i \leqslant n-1$.

Thus, for filiform non-Lie algebras we obtain $D - R_a = \lambda E_{n,2}$, $\lambda \in \mathbb{C}$. □

Remark 2.1. *Let L be a filiform non-Lie Leibniz algebra. If at least one of $\alpha_{i_0} \neq 0$ and $\beta_{j_0} \neq 0$, $i_0, j_0 \in \{4, 5, \ldots, n-1\}$, then we get $AID(L) = Inner(L)$.*

Theorem 2.2. *Let L be an n-dimensional filiform non-Lie Leibniz algebra $F_1(0, \ldots, 0, \alpha_n, \theta)$ or $F_2(0, \ldots, 0, \beta_n, \theta)$. Then at run $\theta = 0$, $\alpha_n \neq 0$ and $\beta_n = 0$, $\gamma \neq 0$ respectively we obtain*

$$AID(L) = Inner(L) \oplus \langle E_{n,2} \rangle,$$

where $E_{n,2}$ is the matrix of the elements in which in the place $(n, 2)$ we have 1, and other elements are 0.

Proof. Let $L = F_1(0, \ldots, 0, \alpha_n, \theta)$. We have to show that $E_{n,2}$ is an almost inner derivation of the algebra L. We take the element $x = \sum\limits_{i=1}^{n} x_i e_i \in L$, then there is $c_x = c_1e_1 + c_2e_2 \in L$ and we check up the following

$$\begin{aligned} E_{n,2}(x) = [x, c_x] &= \left[\sum_{i=1}^{n} x_i e_i, c_1e_1 + c_2e_2\right] = \\ &= c_1(x_1 + x_2)e_3 + c_1x_3e_4 + c_1x_4e_5 + \cdots + c_1x_{n-2}e_{n-1} + (c_1x_{n-1} + c_2(x_1\theta + x_2\alpha_n))e_n. \end{aligned}$$

If $\theta \neq 0$ and $x_3 \neq 0$, then for $x_1 = -\dfrac{x_2\alpha_n}{\theta}$ the map $E_{n,2}$ is not almost inner derivation.

Therefore $\theta = 0$ and for any $x \in L$ choosing $c_1 = 0$, $c_2 = \dfrac{1}{\alpha_n}$ we have $E_{n,2}(x) = x_2e_n$. Hence $E_{n,2} \in AID(L)$.

Let $L = F_2(0, 0, \ldots, 0, \beta_n, \gamma)$. Let $\forall x = \sum_{i=1}^{n} x_i e_i \in L$, then $\exists c_x = c_1 e_1 + c_2 e_2 \in L$ and we obtain the following

$$\begin{aligned} E_{n,2}(x) &= [x, c_x] = \left[\sum_{i=1}^{n} x_i e_i, c_1 e_1 + c_2 e_2\right] = \\ &= c_1 x_1 e_3 + c_1 x_3 e_4 + c_1 x_4 e_5 + \cdots + c_1 x_{n-2} e_{n-1} + (c_1 x_{n-1} + c_2(x_1\beta_n + x_2\gamma))e_n. \end{aligned}$$

If $\beta_n \neq 0$ and $x_4 \neq 0$, then for $x_1 = -\dfrac{x_2\gamma}{\beta_n}$ the derivation $E_{n,2}$ is not almost inner derivation.

Therefore $\beta_n = 0$ and for any $x \in L$ choosing $c_1 = 0,\ c_2 = \dfrac{1}{\gamma}$ we have $E_{n,2}(x) = x_2 e_n$. Hence $E_{n,2} \in AID(L)$. □

Theorem 2.1 and 2.2 imply the following consequence:

Corollary 2.1. *In filiform non-Lie Leibniz algebras, if all parameters are equal to zero, then these algebras turn into a graded algebra. Then the almost inner derivations of graded non-Lie Leibniz algebras coincide with the inner derivations.*

2.4. Almost inner derivations of sme filiform Leibniz algebras

We consider filiform Leibniz algebra $L = F_3(\theta_1, \theta_2, \theta_3)$, which contain filiform Lie algebra [10]:

$$F_3(\theta_1, \theta_2, \theta_3): \begin{cases} [e_1, e_1] = \theta_1 e_n, \quad [e_1, e_2] = -e_3 + \theta_2 e_n, \quad [e_2, e_2] = \theta_3 e_n, \\ [e_i, e_1] = e_{i+1}, \quad 2 \leqslant i \leqslant n-1, \\ [e_1, e_i] = -e_{i+1}, \quad 3 \leqslant i \leqslant n-1, \\ [e_i, e_2] = -[e_2, e_i] = \sum\limits_{k=5}^{n-i+3} \beta_k e_{k+i-3}, \quad 3 \leqslant i \leqslant n-2, \\ [e_i, e_j] = -[e_j, e_i] = 0, \quad i, j \geqslant 3. \end{cases}$$

Theorem 2.3. *Let $L = F_3(\theta_1, \theta_2, \theta_3)$ and let $D \in AID(L)$. Then there exist an element $x \in L$ and $\lambda \in \mathbb{C}$ such that*

$$D - R_x = \lambda E_{n,2}.$$

Proof. Let $L = F_3(\theta_1, \theta_2, \theta_3)$. Let $D \in AID(L)$. Then D induces an almost inner derivation of $\bar{D}$ by $L/\langle e_n \rangle$. By induction, we can assume that after changing D to inner derivation, we have $\bar{D} = \mu E_{n-1,2}$ for some $\mu \in \mathbb{C}$. This implies such that $D(e_1) = \alpha e_n$ for some $\alpha \in \mathbb{C}$. Now we replace D with $D' = D + R_{\alpha e_{n-1}}$. Then we have

$$\begin{aligned} D'(e_1) &= D(e_1) + R_{\alpha e_{n-1}}(e_1) = \alpha e_n + [e_1, \alpha e_{n-1}] = 0, \\ D'(e_i) &= D(e_i) + [e_i, \alpha e_{n-1}] = D(e_i),\ i \geqslant 2. \end{aligned}$$

We get

$$D'(e_2) = D(e_2) = \mu e_{n-1} + \lambda e_n, \quad \mu, \lambda \in \mathbb{C}.$$

Hence, we have the following

$$\begin{aligned} D'(e_3) &= D'\big([e_2, e_1]\big) = [D'(e_2), e_1] = [\mu e_{n-1} + \lambda e_n, e_1] = \mu e_n, \\ D'(e_4) &= D'\big([e_3, e_1]\big) = [D'(e_3), e_1] = [\mu e_n, e_1] = 0, \end{aligned}$$

moreover, $D'(e_i) = 0,\ i \geqslant 5$.

Since we have $D'(e_3) = \mu e_n$ and $D' \in AID(L)$, then there exists an element $a_{e_3} = a_{3,1}e_1 + a_{3,2}e_2 \in L$ such that $D'(e_3) = [e_3, a_{e_3}] = \mu e_n$. Therefore we get the following

$$\mu e_n = [e_3, a_{3,1}e_1 + a_{3,2}e_2] = a_{3,1}e_4 + a_{3,2}(\beta_5 e_5 + \beta_6 e_6 + \cdots + \beta_n e_n).$$

We obtain

$$a_{3,1} = 0, \quad a_{3,2}\beta_i = 0, \quad 5 \leqslant i \leqslant n-1, \quad a_{3,2}\beta_n = \mu.$$

Since we assume $\mu \neq 0$, then we have

$$\beta_i = 0, \quad 5 \leqslant i \leqslant n-1.$$

Now we consider the following

$$D'(e_2) = [e_2, a_{e_2}] = \left[e_2, \sum_{j=1}^{n} a_{2,j}e_j\right] = a_{2,1}e_3 + (a_{2,2}\theta_3 - a_{2,3}\beta_n)e_n.$$

On the other hand $D'(e_2) = \mu e_{n-1} + \lambda e_n$. We have

$$a_{2,1}e_3 + (a_{2,2}\theta_3 - a_{2,3}\beta_n)e_n = \mu e_{n-1} + \lambda e_n.$$

Since we assume that $\mu \neq 0$, this equation does not have a solution, which is a contradiction. Hence indeed $\mu = 0$, and therefore $D' = \lambda E_{2,n}$. □

Proposition 2.5. *Let L be an n-dimensional filiform Leibniz algebra $F_3(\theta_1, \theta_2, \theta_3)$. Then*

$$AID(F_3(\theta_1, \theta_2, \theta_3)) = Inner(F_3(\theta_1, \theta_2, \theta_3)) \oplus \langle E_{n,2} \rangle,$$

where $E_{n,2}$ is the matrix of the elements in which in the place $(n, 2)$ we have 1, and other elements are 0.

Proof. The proof is analogous to Proposition 7.4 in [9]. □

2.5. Almost inner derivations of low dimensional nilpotent Leibniz algebras

N. Jacobson proved the following theorem [12]:

Theorem 2.4. *Every nilpotent Lie algebra has a derivation D which is not inner.*

There is a question: Are almost inner derivations of nilpotent Lie algebras outer derivations? And is the converse right? Generally this question is open. We give an example which answers in the positive on this question.

Example 2.1. *We consider 5-dimensional nilpotent Lie algebra in which there exist almost inner derivations which are not inner [9].*

1) $\mathfrak{g}_{5,3}$: $[e_1, e_2] = e_4$, $[e_1, e_4] = e_5$, $[e_2, e_3] = e_5$, the omitted products are equal to zero. Derivations, inner derivations and almost inner derivations of this algebra have the following matrix forms respectively:

$$Der(\mathfrak{g}_{5,3}) = \begin{pmatrix} a_{1,1} & 0 & 0 & 0 & 0 \\ a_{1,2} & a_{2,2} & 0 & 0 & 0 \\ a_{1,3} & a_{2,3} & 2a_{1,1} & 0 & 0 \\ a_{1,4} & a_{2,4} & -a_{2,2} & a_{1,1} + a_{2,2} & 0 \\ a_{1,5} & a_{2,5} & a_{3,5} & -a_{1,3} + a_{2,4} & 2a_{1,1} + a_{2,2} \end{pmatrix},$$

$$Inner(\mathfrak{g}_{5,3}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \mu_2 & -\mu_1 & 0 & 0 & 0 \\ \mu_4 & \mu_3 & -\mu_2 & -\mu_1 & 0 \end{pmatrix}, \; AID(\mathfrak{g}_{5,3}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ a_{1,4} & a_{2,4} & 0 & 0 & 0 \\ a_{1,5} & a_{2,5} & a_{3,5} & a_{2,4} & 0 \end{pmatrix}.$$

If $a_{1,4} = a_{1,5} = a_{2,4} = a_{2,5} = 0$, then we obtain the matrix of outer derivation of algebra $\mathfrak{g}_{5,3}$:

$$Outer(\mathfrak{g}_{5,3}) = \begin{pmatrix} a_{1,1} & 0 & 0 & 0 & 0 \\ a_{1,2} & a_{2,2} & 0 & 0 & 0 \\ a_{1,3} & a_{2,3} & 2a_{1,1} & 0 & 0 \\ 0 & 0 & -a_{2,2} & a_{1,1}+a_{2,2} & 0 \\ 0 & 0 & a_{3,5} & -a_{1,3} & 2a_{1,1}+a_{2,2} \end{pmatrix}.$$

Therefore, $AID(\mathfrak{g}_{5,3}) \subseteq Outer(\mathfrak{g}_{5,3})$ and any almost inner derivation of the algebra $\mathfrak{g}_{5,3}$ is outer. If in $Outer(\mathfrak{g}_{5,3})$ we have $a_{1,1} = a_{1,2} = a_{1,3} = a_{2,2} = a_{2,3} = 0$, then the space of all outer derivations coincides with the space of all almost inner derivations.

Now we give examples for low dimensional nilpotent Leibniz algebras.

Example 2.2. *Let L be the three-dimensional nilpotent Leibniz algebra:*

$$\begin{aligned}
&L_1(\alpha): && [e_2,e_2]=e_1, && [e_3,e_3]=\alpha e_1, && [e_2,e_3]=e_1, && \alpha\in\mathbb{C},\\
&L_2: && [e_2,e_2]=e_1, && [e_3,e_2]=e_1, && [e_2,e_3]=e_1,\\
&L_3: && [e_2,e_2]=e_1, && [e_3,e_3]=e_1, && [e_3,e_2]=e_1, && [e_2,e_3]=e_1,\\
&L_4: && [e_3,e_3]=e_1,\\
&L_5: && [e_2,e_3]=e_1, && [e_3,e_3]=e_1,\\
&L_6: && [e_3,e_3]=e_1, && [e_1,e_3]=e_2.
\end{aligned}$$

For three-dimensional nilpotent Leibniz algebras L, the following equality

$$AID(L) = Inner(L)$$

holds.

Example 2.3. *Let L be four-dimensional nilpotent Leibniz algebra. Then from [1] there are 28 algebras and we give only those algebras which will be necessary to us:*

$$\begin{aligned}
&L_4: && [e_1,e_1]=e_3, && [e_1,e_2]=\alpha e_4, && [e_2,e_1]=e_3, && [e_2,e_2]=e_4,\\
& && [e_3,e_1]=e_4, && \alpha\in\{0,1\};\\
&L_9: && [e_1,e_1]=e_4, && [e_2,e_1]=e_3, && [e_2,e_2]=e_4, && [e_1,e_2]=-e_3+2e_4,\\
& && [e_3,e_1]=e_4, && [e_1,e_3]=-e_4,\\
&L_{10}: && [e_1,e_1]=e_4, && [e_2,e_1]=e_3, && [e_2,e_2]=e_4, && [e_3,e_1]=e_4,\\
& && [e_1,e_2]=-e_3, && [e_1,e_3]=-e_4;\\
&L_{11}: && [e_1,e_1]=e_4, && [e_1,e_2]=e_3, && [e_2,e_1]=-e_3, && [e_2,e_2]=-2e_3+e_4;\\
&L_{12}: && [e_1,e_1]=e_3, && [e_2,e_1]=e_4, && [e_2,e_2]=-e_3;\\
&L_{13}: && [e_1,e_1]=e_3, && [e_1,e_2]=e_4, && [e_2,e_1]=-\alpha e_3, && [e_2,e_2]=-e_4;\\
&L_{20}: && [e_1,e_2]=e_4, && [e_2,e_1]=\frac{1+\alpha}{1-\alpha}e_4, && [e_2,e_2]=e_3, && \alpha\in\mathbb{C}\setminus\{1\}.
\end{aligned}$$

Let us show the calculation of the dimension of almost inner derivations and the inner derivations of these algebras.

• The algebra L_4 is a filiform algebra from the class $F_1(0,\dots,0,\alpha_n,\theta)$. Therefore, by Theorem 2.2 we have: if $\alpha = 0$, then $AID(L_4) = Inner(L_4)$, and if $\alpha = 1$, then $AID(L_4) = Inner(L_4) \oplus \langle E_{4,2}\rangle$.

• We consider the algebra L_9. Let $D \in AID(L_9)$, then by definition AID for $1 \leqslant i \leqslant 4$ for each e_i there is $a_{e_i} = \sum\limits_{j=1}^{4} a_{i,j}e_j$ and we have the following:

$$D(e_1) = [e_1, a_{e_1}] = -a_{1,2}e_3 + (a_{1,1} + 2a_{1,2} - a_{1,3})e_4, \quad D(e_2) = [e_2, a_{e_2}] = a_{2,1}e_3 + a_{2,2}e_4,$$
$$D(e_3) = [e_3, a_{e_3}] = a_{3,1}e_4, \qquad D(e_4) = [e_4, a_{e_4}] = 0.$$

Since D is derivation, we check the following:

$$a_{3,1}e_4 = D(e_3) = D([e_2, e_1]) = [D(e_2), e_1] + [e_2, D(e_1)] = a_{2,1}e_4,$$

from here we get $a_{2,1} = a_{3,1}$. Therefore, the matrix AID of this algebra has the following form:

$$AID(L_9) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -a_{1,2} & a_{2,1} & 0 & 0 \\ a_{1,1} + 2a_{1,2} - a_{1,3} & a_{2,2} & a_{2,1} & 0 \end{pmatrix},$$

hence $dim AID(L_9) = 4$.

Now we calculate the dimension of the space of inner derivations. To do this, we take the element $x = \sum\limits_{i=1}^{4} x_i e_i$ and consider $R_x(e_i),\ (1 \leqslant i \leqslant 4)$:

$$R_x(e_1) = [e_1, x] = -x_2e_3 + (x_1 + 2x_2 - x_3)e_4, \quad R_x(e_2) = [e_2, x] = x_1e_3 + x_2e_4,$$
$$R_x(e_3) = [e_3, x] = x_1e_4, \qquad R_x(e_4) = [e_4, x] = 0.$$

The matrix of inner derivation of algebra L_9:

$$Inner(L_9) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -x_2 & x_1 & 0 & 0 \\ x_1 + 2x_2 - x_3 & x_2 & x_1 & 0 \end{pmatrix},$$

hence $dim Inner(L_9) = 3$.

From the matrices $AID(L_9)$ and $Inner(L_9)$ it is clear that $AID(L_9) = Inner(L_9) \oplus \langle E_{4,2}\rangle$. Now let's calculate the dimension of $RCAID(L_9)$, for this we take every element of $x = \sum\limits_{i=4}^{4} x_i e_i \in L_9$ and

$$(D - R_x)(x) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -a_{1,2} - x_2 & a_{2,1} - x_1 & 0 & 0 \\ a'_{1,3} - x'_3 & a_{2,2} - x_2 & a_{2,1} - x_1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \bar{0}.$$

Then we have $a_{1,2} = x_2,\ a'_{1,3} = x'_3,\ a_{2,1} = x_1,\ a_{2,2} = x_2$. Hence, $dim RCAID(L_9) = 3$.

• For algebras $L_{10},\ L_{11},\ L_{12},\ L_{20}$ similarly conducted reasoning and calculated dimension $AID(L)$ and $Inner(L)$.

• Now we consider L_{13} and get the following matrices:

$$AID(L_{13}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ a_{1,1} & -a_{2,1} & 0 & 0 \\ a_{1,2} & -a_{2,2} & 0 & 0 \end{pmatrix}, \quad Inner(L_{13}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ x_1 & -x_1 & 0 & 0 \\ x_2 & -x_2 & 0 & 0 \end{pmatrix}.$$

This shows that $dimAID(L_{13}) = 4$, $dimRCAID(L_{13}) = dimInner(L_{13}) = 2$, hence we obtain $AID(L_{13}) = Inner(L_{13}) \oplus \langle E_{3,2} + E_{4,2} \rangle$.

For other algebras, except those shown, almost inner derivations coincide with inner derivations.

Therefore, we have the following table:

Algebra	dim Inner(L)	dimRCAID(L)	dimAID(L)	dimDer(L)	D
L_4	2	2	3	4	$E_{4,2}$
L_9	3	3	4	4	$E_{4,2}$
L_{10}	3	3	4	4	$E_{4,2}$
L_{11}	2	2	3	5	$E_{4,2}$
L_{12}	2	2	3	5	$E_{4,2}$
L_{13}	2	2	4	5	$E_{4,2} + E_{3,2}$
L_{20}	2	2	3	7	$E_{4,2}$

Example 2.4. *Let L be a complex Leibniz algebra of dimension $n \leqslant 2$. Then we have*

$$AID(L) = RCAID(L) = Inner(L).$$

It is clear that for abelian Leibniz algebras $Inner(L) = RCAID(L) = AID(L) = 0$.

References

[1] S.Albeverio, B.A.Omirov, I.S.Rakhimov, Classification of 4-dimensional nilpotent complex Leibniz algebras, *Extracta Mathematicae*, , **21**(2006), no. 3, 197–210.

[2] Sh.A.Ayupov, K.K.Kudaybergenov, Local derivation on finite-dimensional, *Linear Algebra and its Applications*, **493**(2016), 381–398.

[3] Ayupov Sh.A., Kudaybergenov K.K., Local automorphisms on finite-dimensional Lie and Leibniz algebras, 2018, arxiv: 1803.03142v2.

[4] Sh.A.Ayupov, K.K.Kudaybergenov, I.S.Rakhimov, 2-Local derivations on finite-dimensional Lie algebras, *Linear Algebra and its Applications,* **474**(2015), 1–11.

[5] Sh.A.Ayupov, B.A.Omirov, On Leibniz algebras, In: Algebra and Operator Theory (Tashkent, 1997), Kluwer Acad. Publ., Dordrecht, 1998, 1–12.
DOI: 10.1007/978-94-011-5072-9_1

[6] Sh.A.Ayupov, B.A.Omirov, On 3-dimensional Leibniz algebras, *Uzbek Math. Journal*, 1999, 9–14

[7] Sh.A.Ayupov, B.A.Omirov, On some classes of nilpotent Leibniz algebras, *Siberian Math. J.*, **42**(2001), no. 1, 15–24.

[8] M.Breśar, P.Śemrl, Mapping which preserve idempotents, local automorphisms, and local derivations, *Canad. J. Math.*, **45**(1993), 483–496. DOI: 10.4153/CJM-1993-025-4

[9] D.Burde, K.Dekimpe, B.Verbeke, Almost inner derivation of Lie algebras, *Journal of Algebra and Its Applications,* (2018), ID: 119142860. DOI: 10.1142/S0219498818502146.

[10] F.Bratzlavsky, Classification des algébres de Lie de dimension n, de classe $n-1$, dont l'idéal dérivé est commutativ, *Bull. Cl. Sci. Bruxelles*, **60**(1974), 858–865.

[11] N.Jacobson, Lie algebras, Interscience Publishers, Wiley, New York, 1962 .

[12] N.Jacobson, A Note on Automorphisms and Derivations of Lie Algebras, In: Nathan Jacobson Collected Mathematical Papers, Contemporary Mathematicians, Birkhäuser Boston, 1989.

[13] C.S.Gordon, E.N.Wilson, Isospectral defermations of compact solvmanifolds, *J. Differential Geom.*, **19**(1984), no. 1, 214–256.

[14] J.E.Humphreys, Introduction to Lie algebras and Representation theory, Springer-Verlag, New York, 1972.

[15] B.E.Johnson, Local derivations on C*-algebras are derivations, *Trans. Amer. Math. Soc.*, **353**(2001), 313–325.

[16] R.V.Kadison, Local derivations, *J. Algebra*, **130**(1990), 494–509.

[17] A.K.Khudoyberdiev, M.Ladra, B.A.Omirov, The classification of non-characteristically nilpotent filiform Leibniz algebras, *Algebras and Representation Theory*, **17**(2014), no. 3, 945–969.

[18] D.R.Larson, A.R.Sourour, Local derivations and local automorphisms of $B(X)$, Proc. Sympos. Pure Math. 51, Part 2, Providence, Rhode Island, 1990, 187–194.

Почти внутренние дифференцирования некоторых нильпотентных алгебр Лейбница

Жобир К. Адашев
Институт математики АН РУз
Ташкент, Узбекистан
Тууелбай К. Курбанбаев
Каракалпакский государственный университет
Нукус, Узбекистан

Аннотация. В статье исследуется почти внутренние дифференцирования некоторых конечномерных нильпотентных алгебр Лейбница. Мы показываем существование почти внутренних дифференцирований филиформных нелиевых алгебр Лейбница, отличных от внутренних дифференцирований, а также показываем, что почти внутренние дифференцирования некоторых филиформных алгебр Лейбница, содержащих филиформные алгебры Ли, не совпадают с внутренними дифференцированиями.

Ключевые слова: алгебра Лейбница, дифференцирование, внутреннее дифференцирование, почти дифференцирование.

DOI: 10.17516/1997-1397-2020-13-6-746-754
УДК 519.87

Control of Stochastic Processes that Proceeds in the Limited Area

Alexander V. Medvedev
Eugene D. Mikhov*
Siberian Federal University
Krasnoyarsk, Russian Federation

Received 16.06.2020, received in revised form 09.07.2020, accepted 06.09.2020

Abstract. Stochastic process control is considered in the paper. New types of processes (H-processes) are described. Input variables are stochastically related in H-processes. The problem of identification and control of H-processes are considered in detail. The modification of the nonparametric dual control algorithm is developed. The proposed algorithm is compared with the PID algorithm. Application of the proposed algorithm for controlling the H-process with several output variables is presented.

Keywords: nonparametric algorithms, H-prosesses, controlling.

Citation: A.V. Medvedev, E.D. Mikhov, Control of Stochastic Processes that Proceeds in the Limited Area, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 746–754.
DOI: 10.17516/1997-1397-2020-13-6-746-754.

Introduction

Control of multidimensional inertialess processes is considered in the paper.

It is assumed that a controlled process has a parametric structure. In other words, in designing a control algorithm differential equation or system of equations that describes the process is known.

Often the structure of the controlled process is not completely known. In this case, before designing a control algorithm, one need to restore the structure of the controlled process.

Restoring the structure of the process is very complicated process. A control algorithm that does not require restoring of the structure of the control process is considered in the paper. In other words only the values of the input variables $\vec{u}$ and output variables $\vec{x}$ are used in the algorithm. It is assumed that some qualitative characteristics of the process are also known, such as inertia and the degree of nonlinearity of the process.

One of the features of considered processes is stochastic dependence between components of the vector of input variables $(\vec{u})$. That is why the process proceeds not in domain $\Omega(\vec{u})$ determined by the vector of input variables but in some subdomain $\Omega^H(x)$. Often the fact that components of input variables are interdependent is unknown. Of course, the type of relationship is also unknown.

Processes with stochastic interdependency of components of the vector of input variables are called H-processes [1].

Processes in which components of the vector of input variables $\vec{u}$ should be supplied in a certain proportion are H-processes.

Multidimensional H-processes are considered in the paper. Multidimensional H-processes include multiple output variables $\vec{x}$. For each component of the vector of output variables

*edmihov@mail.ru

$x_j, j = \overline{1,k}$ there is its own $\Omega^H(x_j)$, where k is the number of elements in the vector of output variables. In other words, the interrelation between input variables is different for each output variable x_j.

For example, consider some chemical process with two output products (k = 2). To obtain the first output product one should satisfy some conditions. In other words, temperature, pressure, oxygen supply, etc. must be taken into account (components of vector $\vec{u}$). The process proceeds when values of some input variables $\vec{u}$ satisfy some relationships. The domain where these relationships are satisfied is $\Omega^H(x_1)$.

To obtain the second output product one should satisfy some other conditions ($\Omega^H(x_2)$). However, these conditions may be different from conditions for the first product ($\Omega^H(x_1) \neq \Omega^H(x_2)$). Only the values of input variables that satisfy at the same time the conditions for both products allow one to obtain both products. Both products can be obtained in domain $\Omega^H(x_{12}) = \Omega^H(x_1) \cap \Omega^H(x_2)$.

The described above process is an example of a multidimensional H-process.

Obviously, there are many processes with this feature. Standard control algorithms (P -, PI -, PID - regulators) do not use the process domain $\Omega^H(x)$. That is why these control algorithms are not suitable for the processes under consideration.

Thus the need to construct new control algorithms for multidimensional H-processes is actual issue.

The simplified schematics of the considered control loop is shown in Fig. 1, where A is the considered process, $\vec{u}$ is the vector of input variables, $\vec{x}$ is the vector of output variables, $\vec{x}^*$ is the setting action and ξ is the noise.

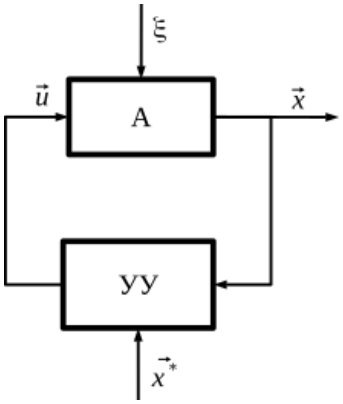


Fig. 1. Schematic representation of the control loop

The H-process does not proceed in domain $\Omega(x)$ but in some subdomain of $\Omega^H(x)$. A schematic representation of the multidimensional H-process is shown in Fig. 2, where $\vec{u}$ is the input vector of dimension n.

The process is characterized by vector of output variables $\vec{x}$ of dimension k. Arrows indicate the interrelation between the input variables.

It is important that the values of the components of the setting action belong to $\Omega^H(x_j)$, $j = \overline{1,k}$. There is no information on interrelation between input variables. Then it is difficult to determine the domain to which every output variable belongs.

Suppose that the process proceeds in domain $\Omega^H(x)$. One should find subdomain $\Omega^{H'}(x) \in \Omega^H(x)$ which is part of process domain (Fig. 3).

However, it can be difficult to find domain $\Omega^{H'}(\vec{u})$ because domain $\Omega^H(\vec{u})$ is unknown. In addition there can be several intersections of domains $\Omega^H(x_j)$, $j = \overline{1,k}$. Then one needs to

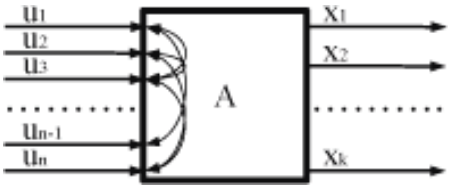


Fig. 2. Schematic representation of multidimensional H-process

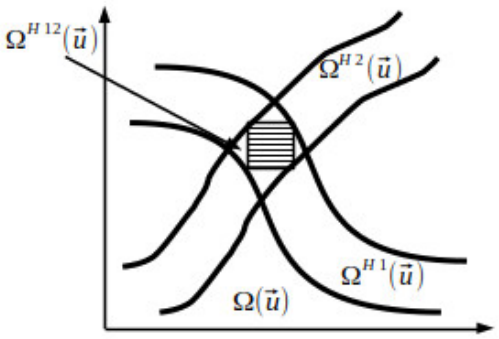


Fig. 3. Isolation of H-process flow domain

decide: what intersections should be used to control the process? Should the transition between these intersections be used in control and how the transition influences the process control?

These difficulties demonstrate that development of control algorithms for multi-dimensional H-process is the topical problem. The control algorithms should use the domain where the process proceeds and it should include some analysis of the control object. In other words, the control algorithm should be the adaptive algorithm.

The nonparametric dual control algorithm [1] is used in the paper. The nonparametric dual control algorithm is based on two methods: the dual control method developed by A. A. Feldbaum [2] and the nonparametric regression function estimation method [4]. The nonparametric dual control algorithm was developed by A.V. Medvedev [3].

The nonparametric regression function estimation

Let us consider statically independent observations of two random variables $(x, y) = = (x_1, y_1), \ldots, (x_n, y_n)$ that are distributed with unknown frequency function $P(x, y)$.

Let us assume that $p(x) > 0\, \forall x \in \Omega(x)$. To approximate the unknown stochastic relationship between y and x the regression is often used [4]:

$$y = f(x) = \left(\int_{\Omega(y)} P(x, y) dy \right)^{-1} \left(y \int_{\Omega(y)} P(x, y) dy \right) \tag{1}$$

Nonparametric estimation of relation (1) is

$$\hat{y} = \hat{f}(x) = \left(\sum_{i=1}^{n} \Phi\left(\frac{x - x_i}{C_n} \right) \right)^{-1} \sum_{i=1}^{n} y_i \Phi\left(C_n^{-1} (x - x_i) \right) \tag{2}$$

When $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ are vectors relation (2) becomes

$$\hat{y}_d = \frac{\sum\limits_{i=1}^{n} y_i^d \prod_{j=k}^{k} \Phi(C_j^{-1}(n)(x^j - x_i^j))}{\sum\limits_{i=1}^{n} \prod_{j=k}^{k} \Phi(C_j^{-1}(n)(x^j - x_i^j))}, \quad d = \overline{1,k} \tag{3}$$

The nonparametric estimate of the regression curve is convergent, i.e.,

$$\lim_{n\to\infty} M((f(x) - f_n(x))^2) = 0, \forall x \in \Omega(x) \tag{4}$$

$$\lim_{n\to\infty} M(f(x)) = f(x), \forall x \in \Omega(x). \tag{5}$$

Information on the parametric structure of the object is not needed for the noparametric estimate of regression function.

Nonparametric dual control

In the case when control algorithm includes control and investigation of the system, it is called dual control algorithm.

Dual control algorithm was developed by A. A. Feldbaum. The nonparametric dual control algorithm is

$$u_{s+1} = u_s^* + \delta u_{s+1}, \tag{6}$$

where u_s^* is "knowledge" of the object, δu_{s+1} is "learning" search steps (in the classic form of nonparametric dual control algorithm) and δu_{s+1} is

$$\delta u_{s+1} = \xi(x_{s+1}^* - x_s). \tag{7}$$

Using the noparametric estimate of the regression function $(x_i, u_i), i = \overline{1,s}$, we obtain the estimate of the object $\hat{x} = \hat{f}(\vec{u})$ as

$$\hat{x}(u) = \frac{\sum\limits_{i=1}^{n} x_i \Phi(\frac{u-u_i}{c_s})}{\sum\limits_{i=1}^{n} \Phi(\frac{u-u_i}{c_s})}. \tag{8}$$

Here bell-shaped functions $\Phi(\cdot)$ and smooth coefficient c_s satisfy convergence condition, $u = f^{-1}(x)$, where $f^{-1}(x)$ is the inverse of $f(u)$, and u_s^* is

$$u_s^* = \frac{\sum\limits_{i=1}^{n} u_i \Phi\big(\frac{\Phi(x^* - x_i)}{c_s}\big)}{\sum\limits_{i=1}^{n} \frac{\Phi(x^* - x_i)}{c_s}}, \tag{9}$$

where x^* is the setting action.

At the beginning of process control, second component δu_{s+1} is more important component of control. This is the time of active investigating of the dual control system. This stage begins with receiving of the first values of the input and output variables. The first component (u_{s*}) becomes more important component of control after stage of active investigating. Thus, there are stage of object investigation and stage of action in the process of dual control.

Modification of nonparametric dual control algorithm for multidimensional H-processes

In multidimensional H-process control the action cannot be arbitrarily specified as it is considered in control theory. This is due to the fact that it is possible to set a vector of action such that $\prod i = 1^k \Omega_i^H(\vec{x}^*) = \varnothing$. In other words, this action is not achievable for all components of vector $\vec{x}^*$ at the same time. That is why it is important to set $\vec{x}^* \in \prod_{i=1}^{k} \Omega_i^H(\vec{x}^*)$, i.e., define $x_1^*, x_2^*, \ldots, x_k^*$.

We propose the following method:

1. Calculate the value $\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\Big(\frac{x_j^* - x_{ij}}{c_{sj}}\Big)$, where $\vec{x}^*$ is the action and s is the size of sample observations.

2. If the calculated value $\sum_{i=1}^{s} \prod_{j=1}^{k} \Phi\Big(\frac{x_j^* - x_{ij}}{c_{sj}}\Big)$ is not equal to zero then the action is achievable otherwise the action may not be achieved.

Let us note that in nonparametric dual control the calculation of the search step $\delta\vec{u}_{s+1}$ is performed with the use of (7). In the case of H-process with several output variables the described method for calculating the search step $\delta\vec{u}_{s+1}$ is not suitable because the input action must belong to $\prod_{j=1}^{k} \Omega^H(\vec{u})$.

Taking this into account, we propose to use an algorithm with punishment to calculate $\delta\vec{u}_{s+1}$. This method determines the reachability of the action and it can be described as follows

1) $\vec{u}_s(\vec{x}^*)$ is calculated;

2) a random vector $\delta\vec{u}_{s+1}$ is generated;

3) $\vec{u}_{s+1}(x^*)$ is calculated;

4) if $\vec{u}_{s+1} \in \Omega^H(\vec{u})$ then $\vec{u}_{s+1}$ is used as a control action otherwise we return to step 2;

5) if $\sum_{i=1}^{k} |x_{i,s+1} - x_i^*| < \sum_{i=1}^{k} |x_{i,s} - x_i^*|$ then $\delta\vec{u}_{s+1}$ is used as the next value of the search step $\delta\vec{u}_{s+2}$ otherwise the random vector $\delta\vec{u}_{s+2}$ is generated;

6) go back to step 1.

The length of the vector $\delta\vec{u}$ is $m|x^* - x_s|$, where m is the preassigned coefficient.

Results of computer simulation

Examples of control of the inertialess process and H-process are presented. Standard regulator and regulator on the basis of the nonparametric algorithm of dual control are used in simulations.

Firstly, the PID regulator is compared with the modified algorithm of nonparametric dual control. Secondly, the nonparametric dual control algorithm is used to control the multidimensional H-process.

The simulated multidimensional process is defined as follows

$$\begin{aligned} x_1(\vec{u}) &= f(u_1, u_2) = u_1 * 3 * sin(u_2) + 7 + \xi_1, \\ x_2(\vec{u}) &= f(u_1, u_2) = 4 * (u_1) + u_2 + \xi_2. \end{aligned} \tag{10}$$

Let us note that structure of the object is not used in the algorithm. It is used only to generate a sample of observations.

Next the result of control based on the PID regulator is presented.

The result of control over x_1 is shown in the upper part of Fig. 4. The result of control over x_2 is shown in the lower part of the figure. The value of the action is marked by the dashed line. The iteration number is shown on the abscissa.

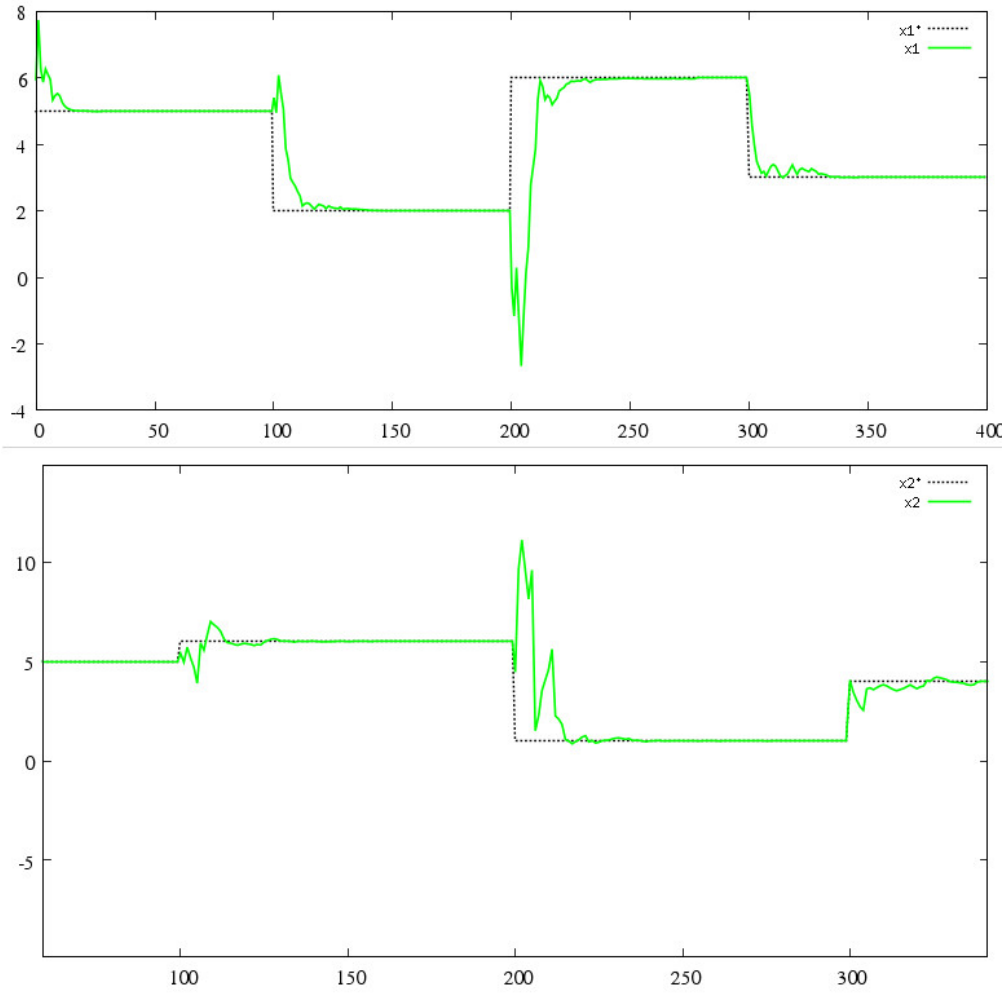


Fig. 4. Result of the PID control

Results presented in Fig. 4 demonstrate that PID algorithm successfully controls object (10). This algorithm is not adaptive, and it does not include training. This means that algorithm controls the process without using a sample of observations to improve its characteristics. That is why the control efficiency did not increase.

Nonparametric dual control algorithm is used to control process (10) in the next experiment.

Unlike the PID algorithm used in the previous experiment (Fig. 5) this algorithm includes training. This is confirmed by the fact that after training the control is more efficient.

Next we consider a multidimensional inertialess H-process. There is a stochastic relationship between input variables in the H-process.

The process is described by the system of equations (10). The relationship between input

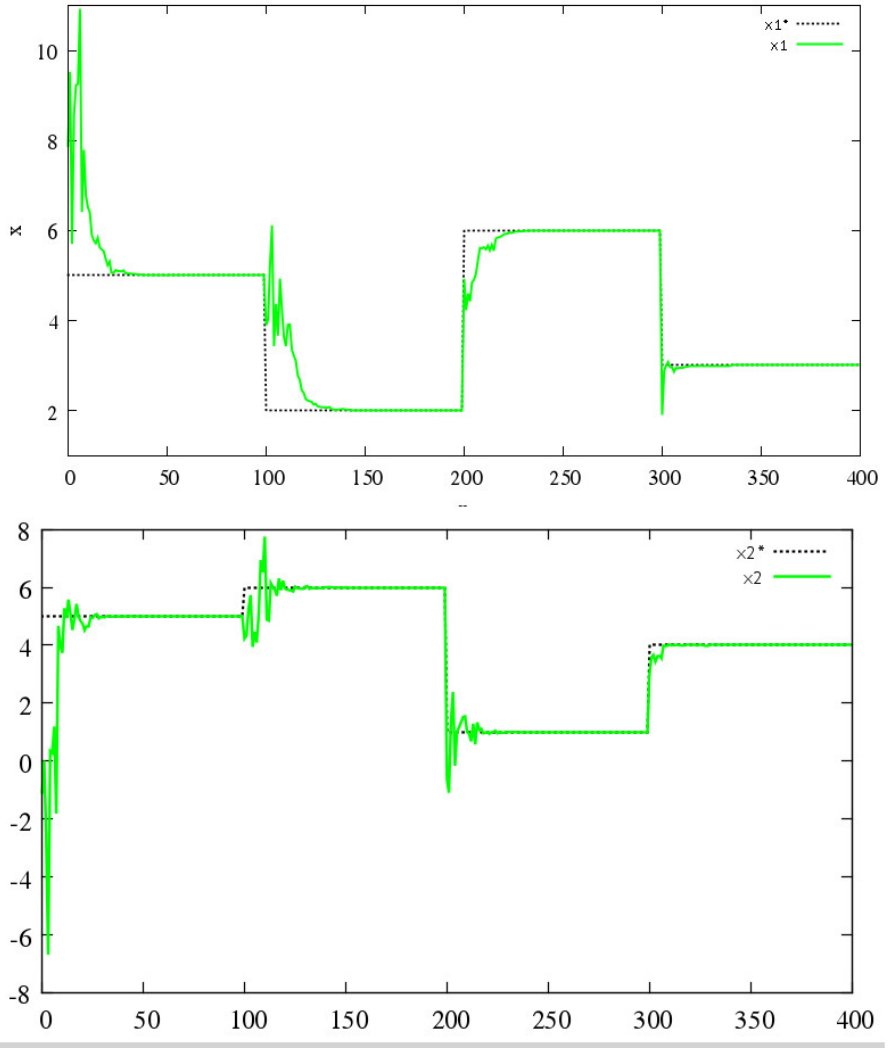


Fig. 5. Result of nonparametric dual control

variables is described by the following equations

$$u_2 = u_1 * 5 - 3 + \xi_1, \xi_1 \in (-0.4; 0.4), \tag{11}$$

$$u_1 = u_2 * 3 + 1 + \xi_2, \xi_2 \in (-0.6; 0.6). \tag{12}$$

Variables ξ_1, ξ_2 characterize the "width" of domains $\Omega^{H1}(\vec{u})$ and $\Omega^{H2}(\vec{u})$, respectively.

It is difficult to control this H-process using the PID algorithm because this algorithm does not take into account domain ($\Omega^{H1}(\vec{u})$, $\Omega^{H2}(\vec{u})$). This is important in the case of the H-process.

The H-process under consideration is controlled with the use of the modified nonparametric dual control algorithm (Fig. 6).

The modification of the nonparametric dual control algorithm can be successfully applied to control the multidimensional H-process. It is demonstrated in Fig. 6. The nonparametric dual control algorithm is adaptive.

Therefore, after training, the proposed modification more effectively controls the process.

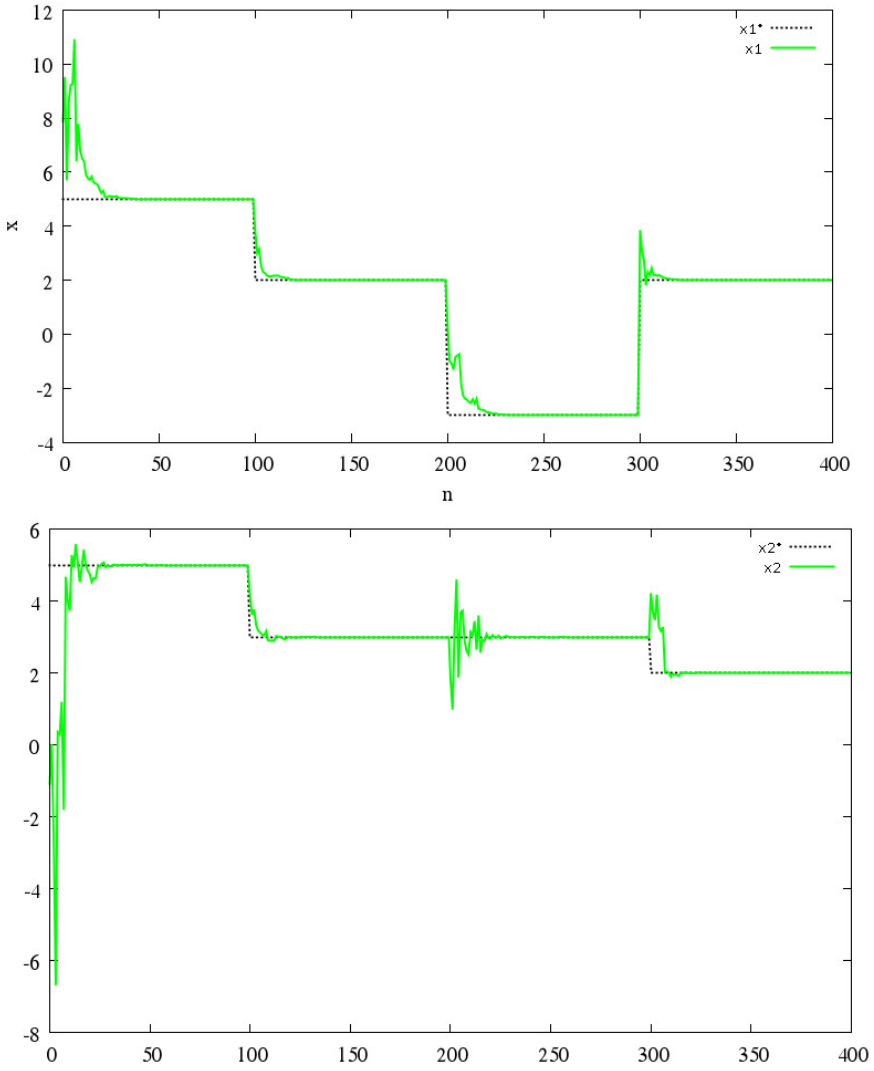


Fig. 6. Result of control with the modified nonparametric dual control algorithm

Conclusion

A modification of the nonparametric dual control algorithm was proposed. A feature of the proposed modification is a new choice of the search step which takes into account the domain of the process. The proposed modification was applied to control multidimensional inertialess processes with interdependent input variables.

The modified nonparametric dual control algorithm and the PID algorithm was compared. It was demonstrated that the modified nonparametric dual control algorithm is adaptive. After training this algorithm controls the process more effectively than the PID algorithm. The modification of nonparametric dual control was applied to a multidimensional inertialess H-process. The proposed algorithm successfully controls the multidimensional inertialess H-process.

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation under the project "Creation of a production of earth stations of advanced

satellite communications systems to ensure the coherence of hard, northern and Arctic territory of Russian Federation", implemented with the participation of the Siberian Federal University (agreement number 075 -11-2019-078 dated 13.12.2019).

References

[1] A.V.Medvedev, Nonparametric adaptation systems, Novosibirsk, Nauka, 1983 (in Russian).

[2] A.A.Feldbaum, Fundamentals of the theory of optimal automatic systems, Moscow, Fizmatgiz, 1963 (in Russian).

[3] A.Medvedev, Theory of nonparametric systems. Modeling, *Bulletin of SibSAU*, **30**(2010), no. 4, 4–9 (in Russian).

[4] P.Eykhoff, Fundamentals of identification of control systems, Moscow, Mir, 1975.

Управление стохастическими процессами с ограниченной областью протекания

Александр В. Медведев
Евгений Д. Михов
Сибирский федеральный университет
Красноярск, Российская Федерация

Аннотация. В статье рассматриваются вопросы управления стохастическими процессами. Описаны новые виды процессов (Н-процессы), в которых имеется стохастическая зависимость между входными переменными. Для решения проблем, возникающих при решении задачи идентификации и управления, была предложена модификация алгоритма непараметрического дуального управления. Проведены эксперименты, в которых предложенная модификация алгоритма сравнивается по эффективности с ПИД-регулятором. В конце статьи представлен эксперимент по управлению Н-процессом с несколькими выходными переменными при помощи разработанного алгоритма.

Ключевые слова: непараметрические алгоритмы, Н-процессы, управление.

DOI: 10.17516/1997-1397-2020-13-6-755-762
УДК 517.95

Mixed Biharmonic Dirichlet-Neumann Problem in Exterior Domains

Hovik A. Matevossian*
Federal Research Center "Computer Science and Control" RAS
Moscow, Russian Federation
Moscow Aviation Institute (National Research University)
Moscow, Russian Federation

Received 17.09.2019, received in revised form 04.06.2020, accepted 17.10.2020

Abstract. We study the unique solvability of the mixed Dirichlet-Neumann problem for the biharmonic equation in the exterior of a compact set under the assumption that solutions of this problem have bounded Dirichlet integrals with the weight $|x|^a$. Depending on the value of the parameter a, we obtained uniqueness (non-uniqueness) theorems of the problem and present exact formulas for the dimension of the space of solutions of the mixed Dirichlet-Neumann problem.

Keywords: biharmonic operator, Dirichlet-Neumann problem, weighted Dirichlet integral.

Citation: H.A. Matevossian, Mixed Biharmonic Dirichlet-Neumann Problem in Exterior Domains, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 755–762. DOI: 10.17516/1997-1397-2020-13-6-755-762.

1. Introduction and preliminaries

Let Ω be an unbounded domain in $\mathbb{R}^n$, $n \geqslant 2$, $\Omega = \mathbb{R}^n \setminus \overline{G}$ with the boundary $\partial\Omega \in C^2$, where G is a bounded simply connected domain (or a union of finitely many such domains) in $\mathbb{R}^n$, $0 \in G$, $\overline{\Omega} = \Omega \cup \partial\Omega$ is the closure of Ω, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ and $|x| = \sqrt{x_1^2 + \dots + x_n^2}$.

In the domain Ω we consider the following mixed problems for the biharmonic equation

$$\Delta^2 u = 0 \tag{1}$$

with the Dirichlet–Neumann boundary conditions

$$u\big|_{\Gamma_1} = \frac{\partial u}{\partial \nu}\Big|_{\Gamma_1} = 0, \quad \Delta u\big|_{\Gamma_2} = \frac{\partial \Delta u}{\partial \nu}\Big|_{\Gamma_2} = 0, \tag{2}$$

where $\overline{\Gamma}_1 \cup \overline{\Gamma}_2 = \partial\Omega$, $\Gamma_1 \cap \Gamma_2 = \emptyset$, $\operatorname{mes}_{n-1} \Gamma_1 \neq 0$, $\nu = (\nu_1, \dots, \nu_n)$ is the outer unit normal vector to $\partial\Omega$.

As is well known, if Ω is an unbounded domain, one should additionally characterize the behavior of the solution at infinity. As a rule, to this end, one usually poses either the condition that the Dirichlet (energy) integral is finite or a condition on the character of vanishing of the modulus of the solution as $|x| \to \infty$. Such conditions at infinity are natural and were studied by several authors (e.g., [6–8]).

Elliptic problems with parameters in the boundary conditions have been called Steklov or Steklov-type problems, since their first appearance in [27]. For the biharmonic operator, these

*hmatevossian@graduate.org https://orcid.org/0000-0002-9895-9628

conditions were first considered in [1, 9] and [25], where the isoperimetric properties of the first eigenvalue were studied.

Note that standard elliptic regularity results are available in [3]. The monograph covers higher order linear and nonlinear elliptic boundary value problems, mainly with the biharmonic or polyharmonic operator as the leading principal part. The underlying models and, in particular, the role of different boundary conditions are explained in detail. As for linear problems, after a brief summary of the existence theory and L^p and Schauder estimates, the focus is on positivity. The required kernel estimates are also presented in detail.

In [2], the boundary value problems for the biharmonic equation and the Stokes system are studied in a half space, and, using the Schwarz reflection principle in weighted L^q-space, the uniqueness of solutions of the Stokes system or the biharmonic equation is proved.

In the present note, this condition is the boundedness of the weighted Dirichlet integral:

$$D_a(u,\Omega) \equiv \int_\Omega |x|^a \sum_{|\alpha|=2} |\partial^\alpha u|^2 \, dx < \infty, \quad a \in \mathbb{R}.$$

In various classes of unbounded domains with finite weighted Dirichlet (energy) integral, one of the author [10–23] studied uniqueness (non–uniqueness) problem and found the dimensions of the spaces of solutions of boundary value problems for the elasticity system and the biharmonic (polyharmonic) equation.

By developing an approach based on the use of Hardy type inequalities [6–8], in the present note, we obtain a uniqueness (non–uniqueness) criterion for a solution of the mixed Dirichlet–Neumann problem for the biharmonic equation.

Notation: $C_0^\infty(\Omega)$ is the space of infinitely differentiable functions in Ω with compact support in Ω. We denote by $H^m(\Omega,\Gamma)$, $\Gamma \subset \overline{\Omega}$, the Sobolev space of functions in Ω obtained by the completion of $C^\infty(\overline{\Omega})$ vanishing in a neighborhood of Γ with respect to the norm

$$||u; H^m(\Omega,\Gamma)|| = \left(\int_\Omega \sum_{|\alpha|\leqslant m} |\partial^\alpha u|^2 dx \right)^{1/2}, \quad m = 1, 2,$$

where $\partial^\alpha \equiv \partial^{|\alpha|}/\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}$, $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index, $\alpha_i \geqslant 0$ are integers, and $|\alpha| = \alpha_1 + \dots + \alpha_n$; if $\Gamma = \emptyset$, we denote $H^m(\Omega,\Gamma)$ by $H^m(\Omega)$.

$\overset{\circ}{H}{}^m(\Omega)$ is the space obtained by the completion of $C_0^\infty(\Omega)$ with respect to the norm $||u; H^m(\Omega)||$.

$\overset{\circ}{H}{}^m_{loc}(\Omega)$ is the space obtained by the completion of $C_0^\infty(\Omega)$ with respect to the family of semi-norms

$$\|u; H^m(\Omega \cap B_0(R))\| = \left(\int\limits_{\Omega \cap B_0(R)} \sum_{|\alpha|\leqslant m} |\partial^\alpha u|^2 \, dx \right)^{1/2}$$

for all open balls $B_0(R) := \{x : |x| < R\}$ in $\mathbb{R}^n$ for which $\Omega \cap B_0(R) \neq \emptyset$.

Let $\binom{n}{k}$ be the (n,k)-binomial coefficient, $\binom{n}{k}$=0 for $k > n$.

2. Definitions and auxiliary statements

Definition 2.1. *A solution of the homogenous biharmonic equation (1) in Ω is a function $u \in H^2_{loc}(\Omega)$ such that for every function $\varphi \in C_0^\infty(\Omega)$, the following integral identity holds:*

$$\int_\Omega \Delta u \, \Delta\varphi \, dx = 0.$$

Lemma 2.2. *Let u be a solution of equation (1) in Ω such that $D_a(u,\Omega)<\infty$. Then*

$$u(x)=P(x)+\sum_{\beta_0<|\alpha|\leqslant\beta}\partial^\alpha\Gamma(x)C_\alpha+u^\beta(x),\quad x\in\Omega, \tag{3}$$

where $P(x)$ is a polynomial, $\operatorname{ord}P(x)<m_0=\max\{2,2-n/2-a/2\}$, $\beta_0=2-n/2+a/2$, *$\Gamma(x)$ is the fundamental solution of equation (1), $C_\alpha=\text{const}$, $\beta\geqslant 0$ is an integer, and the function u^β satisfies the estimate:*

$$|\partial^\gamma u^\beta(x)|\leqslant C_{\gamma\beta}|x|^{3-n-\beta-|\gamma|},\quad C_{\gamma\beta}=\text{const},$$

for every multi-index γ.

Remark 2.3. *As is known [26], the fundamental solution $\Gamma(x)$ of the biharmonic equation has the form*

$$\Gamma(x)=\begin{cases} C|x|^{4-n} & \text{if } 4-n<0 \text{ or } n \text{ is odd},\\ C|x|^{4-n}\ln|x| & \text{if } 4-n\geqslant 0 \text{ and } n \text{ is even}.\end{cases}$$

Proof of Lemma 2.2. Consider the function $v(x)=\theta_N(x)u(x)$, where $\theta_N(x)=\theta(|x|/N),\theta\in C^\infty(\mathbb{R}^n)$, $0\leqslant\theta\leqslant 1$, $\theta(s)=0$ for $s\leqslant 1$, $\theta(s)=1$ for $s\geqslant 2$, while $N\gg 1$ and $G\subset\{x:|x|<N\}$. We extend v to $\mathbb{R}^n$ by setting $v=0$ on $G=\mathbb{R}^n\setminus\overline{\Omega}$.

Then the function v belongs to $C^\infty(\mathbb{R}^n)$ and satisfies the equation

$$\Delta^2 v=f,$$

where $f\in C_0^\infty(\mathbb{R}^n)$ and $\operatorname{supp}f\subset\{x:|x|<2N\}$. It is easy to see that $D_a(v,\mathbb{R}^n)<\infty$.

We can now use Theorem 1 of [5] since it is based on Lemma 2 of [5], which imposes no constraint on the *sign* of σ. Hence, the expansion

$$v(x)=P(x)+\sum_{\beta_0<|\alpha|\leqslant\beta}\partial^\alpha\Gamma(x)C_\alpha+v^\beta(x),$$

holds for each a, where $P(x)$ is a polynomial of order $\operatorname{ord}P(x)<m_0=\max\{2,2-n/2-a/2\}$, $\beta_0=2-n/2+a/2$, $C_\alpha=\text{const}$ and

$$|\partial^\gamma v^\beta(x)|\leqslant C_{\gamma\beta}|x|^{3-n-\beta-|\gamma|},\quad C_{\gamma\beta}=\text{const}\,.$$

Therefore, by the definition of v, we obtain (3). The proof of Lemma 2.2 is complete. □

Definition 2.4. *A function u is a solution of the mixed Dirichlet–Neumann problem (1), (2), if $u\in\overset{\circ}{H}{}^2_{loc}(\Omega,\Gamma_1)$ such that for every function $\varphi\in C_0^\infty(\mathbb{R}^n)$, $\varphi=0$ in the neighborhood of Γ_1, the following integral identity holds:*

$$\int_\Omega \Delta u\,\Delta\varphi\,dx=0. \tag{4}$$

3. Main Results

Theorem 3.1. *The mixed Dirichlet–Neumann problem (1), (2) with the condition $D(u,\Omega)<\infty$ has $n+1$ linearly independent solutions.*

Proof. For any nonzero vector A in $\mathbb{R}^n$, we construct a generalized solution u_A of the biharmonic equation (1) with the boundary conditions

$$u_A(x)\big|_{\Gamma_1}=(Ax)\big|_{\Gamma_1},\quad \frac{\partial u_A(x)}{\partial\nu}\Big|_{\Gamma_1}=\frac{\partial(Ax)}{\partial\nu}\Big|_{\Gamma_1},\quad \Delta u_A\big|_{\Gamma_2}=\frac{\partial\Delta u_A}{\partial\nu}\Big|_{\Gamma_2}=0, \tag{5}$$

and the condition

$$\chi(u_A,\Omega)\equiv\begin{cases}\displaystyle\int_\Omega\left(\frac{|u_A(x)|^2}{|x|^4}+\frac{|\nabla u_A(x)|^2}{|x|^2}+|\nabla\nabla u_A(x)|^2\right)dx<\infty\\ \qquad\qquad\qquad\qquad\qquad\qquad\text{for } n>4,\\ \displaystyle\int_\Omega\left(\frac{|u_A(x)|^2}{||x|^2\ln|x||^2}+\frac{|\nabla u_A(x)|^2}{||x|\ln|x||^2}+|\nabla\nabla u_A(x)|^2\right)dx<\infty\\ \qquad\qquad\qquad\qquad\qquad\qquad\text{for } 2\leqslant n\leqslant 4,\end{cases}\tag{6}$$

for $A,x\in\mathbb{R}^n$, where Ax denotes the standard scalar product of A and x.

Such a solution of problem (1), (5) can be constructed by the variational method [26], minimizing the functional

$$\Phi(v)=\frac{1}{2}\int_\Omega|\Delta v|^2\,dx$$

in the class of admissible functions $\big\{v\colon v\in H^2(\Omega),\, v(x)\big|_{\Gamma_1}=(Ax)\big|_{\Gamma_1},\, \dfrac{\partial v(x)}{\partial\nu}\big|_{\Gamma_1}=\dfrac{\partial(Ax)}{\partial\nu}\big|_{\Gamma_1},$ v is compactly supported in $\overline{\Omega}\big\}$. The validity of condition (6) as a consequence of the Hardy inequality follows from the results in [6–8].

Now, for any arbitrary number $e\neq 0$, we construct a generalized solution u_e of equation (1) with the boundary conditions

$$u_e\big|_{\Gamma_1}=e,\quad \left.\frac{\partial u_e}{\partial\nu}\right|_{\Gamma_1}=0,\quad \Delta u_e\big|_{\Gamma_2}=\left.\frac{\partial\Delta u_e}{\partial\nu}\right|_{\Gamma_2}=0,\tag{7}$$

and the condition

$$\chi(u_e,\Omega)\equiv\begin{cases}\displaystyle\int_\Omega\left(\frac{|u_e(x)|^2}{|x|^4}+\frac{|\nabla u_e(x)|^2}{|x|^2}+|\nabla\nabla u_e(x)|^2\right)dx<\infty\\ \qquad\qquad\qquad\qquad\qquad\qquad\text{for } n>4,\\ \displaystyle\int_\Omega\left(\frac{|u_e(x)|^2}{||x|^2\ln|x||^2}+\frac{|\nabla u_e(x)|^2}{||x|\ln|x||^2}+|\nabla\nabla u_e(x)|^2\right)dx<\infty\\ \qquad\qquad\qquad\qquad\qquad\qquad\text{for } 2\leqslant n\leqslant 4.\end{cases}\tag{8}$$

The solution of problem (1), (7) is also constructed by the variational method with the minimization of the corresponding functional in the class of admissible functions $\{v : v\in H^2(\Omega),\, v\big|_{\Gamma_1}=e,\, \left.\dfrac{\partial v}{\partial\nu}\right|_{\Gamma_1}=0,\quad v$ is compactly supported in $\overline{\Omega}\}$. The condition (8) as a consequence of the Hardy inequality follows from the results in [6–8].

Consider the function $v(x)=(u_A(x)-Ax)-(u_e-e)$. Obviously, v is a solution of problem (1), (2):

$$\Delta^2 v=0,\quad x\in\Omega,\quad v\big|_{\Gamma_1}=\left.\frac{\partial v}{\partial\nu}\right|_{\Gamma_1}=0,\quad \Delta v\big|_{\Gamma_2}=\left.\frac{\partial\Delta v}{\partial\nu}\right|_{\Gamma_2}=0.$$

One can easily see that $v\not\equiv 0$ and $D(v,\Omega)<\infty$.

To each nonzero vector $\mathbf{A}=(A_0,A_1,\dots,A_n)$ in $\mathbb{R}^{n+1}$, there corresponds a nonzero solution $v_{\mathbf{A}}=(v_{A_0},v_{A_1},\dots,v_{A_n})$ of problem (1), (2) with the condition $D(v_{\mathbf{A}},\Omega)<\infty$, and moreover,

$$v_{\mathbf{A}}(x)=u_A(x)-u_e-Ax+e.$$

Let $A_0,A_1,\dots,A_n$ be a basis in $\mathbb{R}^{n+1}$. Let us prove that the corresponding solutions $v_{A_0},v_{A_1},\dots,v_{A_n}$ are linearly independent. Let

$$\sum_{i=0}^{n} C_i v_{A_i} \equiv 0, \qquad C_i = \text{const}\,.$$

Set $W(x) \equiv \sum\limits_{i=1}^{n} C_i A_i x - C_0 e$. We have $W(x) = \sum\limits_{i=1}^{n} C_i u_{A_i}(x) - C_0 u_e$,

$$\int_\Omega |x|^{-2}|\nabla W|^2\,dx < \infty, \quad n > 4; \qquad \int_\Omega ||x|\ln|x||^{-2}|\nabla W|^2\,dx < \infty, \quad 2 \leqslant n \leqslant 4.$$

Let us show that

$$W(x) \equiv \sum_{i=1}^{n} C_i A_i x - C_0 e \equiv 0.$$

Let $T = \sum\limits_{i=0}^{n} C_i A_i = (t_0, \dots, t_n)$, where $A_0 = -e$. Then

$$\int_\Omega |x|^{-2}|\nabla W|^2\,dx = \int_\Omega |x|^{-2}(t_1^2 + \dots + t_n^2)\,dx = \infty, \quad n > 4,$$
$$\int_\Omega ||x|\ln|x||^{-2}|\nabla W|^2\,dx = \int_\Omega ||x|\ln|x||^{-2}(t_1^2 + \dots + t_n^2)\,dx = \infty, \quad 2 \leqslant n \leqslant 4,$$

if $T \neq 0$.

Consequently, $T = \sum\limits_{i=0}^{n} C_i A_i = 0$, and since the vectors $A_0, A_1, \dots, A_n$ are linearly independent, we obtain $C_i = 0,\ i = 0, 1, \dots, n$.

Thus, the Dirichlet–Neumann problem (1), (2) with the condition $D(u, \Omega) < \infty$ has at least $n + 1$ linearly independent solutions.

Let us prove that each solution u of problem (1), (2) with the condition $D(u, \Omega) < \infty$ can be represented as a linear combination of the functions $v_{A_0}, v_{A_1}, \dots, v_{A_n}$, i.e.

$$u = \sum_{i=0}^{n} C_i v_{A_i}, \qquad C_i = \text{const}\,.$$

Since $A_0, A_1, \dots, A_n$ is a basis in $\mathbb{R}^{n+1}$, it follows that there exist constants $C_0, C_1, \dots, C_n$ such that

$$A = \sum_{i=0}^{n} C_i A_i.$$

We set

$$u_0 \equiv u - \sum_{i=0}^{n} C_i v_{A_i}.$$

Obviously, the function u_0 is a solution of problem (1), (2), and $D(u_0, \Omega) < \infty,\ \chi(u_0, \Omega) < \infty$.

Let us show that $u_0 \equiv 0,\ x \in \Omega$. To this end, we substitute the function $\varphi(x) = u_0(x)\theta_N(x)$ into the integral identity (4) for the function u_0, where $\theta_N(x) = \theta(|x|/N),\ \theta \in C^\infty(\mathbb{R}),\ 0 \leqslant \theta \leqslant 1$, $\theta(s) = 0$ for $s \geqslant 2$ and $\theta(s) = 1$ for $s \leqslant 1$; then we obtain

$$\int_\Omega (\Delta u_0)^2 \theta_N(x)\,dx = -J_1(u_0) - J_2(u_0), \tag{9}$$

where

$$J_1(u_0) = 2\int_\Omega \Delta u_0\,\nabla u_0\,\nabla\theta_N(x)\,dx, \qquad J_2(u_0) = \int_\Omega u_0\,\Delta u_0\,\Delta\theta_N(x)\,dx.$$

By applying the Cauchy–Schwarz inequality and by taking into account the conditions $D(u_0,\Omega)<\infty$ and $\chi(u_0,\Omega)<\infty$, one can easily show that $J_1(u_0)\to 0$ and $J_2(u_0)\to 0$ as $N\to\infty$. Consequently, by passing to the limit as $N\to\infty$ in (9), we obtain

$$\int_\Omega (\Delta u_0)^2\,dx=0.$$

Therefore, we have

$$\Delta u_0=0,\quad x\in\Omega,$$

$$u_0\big|_{\Gamma_1}=\left.\frac{\partial u_0}{\partial\nu}\right|_{\Gamma_1}=0,\quad \Delta u_0\big|_{\Gamma_2}=\left.\frac{\partial \Delta u_0}{\partial\nu}\right|_{\Gamma_2}=0.$$

Hence, it follows [4, Ch.2] that $u_0\equiv 0$ in Ω. The proof of the theorem is complete. □

Theorem 3.2. *The mixed Dirichlet–Neumann problem (1), (2) with the condition $D_a(u,\Omega)<\infty$ has:*

(i) the trivial solution for $n-2\leqslant a<\infty$, $n>4$;

(ii) n linearly independent solutions for $n-4\leqslant a<n-2$, $n>4$;

(iii) $n+1$ linearly independent solutions for $-n\leqslant a<n-4$, $n>4$;

(iv) $k(r,n)$ linearly independent solutions for $-2r+2-n\leqslant a<-2r+4-n$, $r>1$, $n>4$, where

$$k(r,n)=\binom{r+n}{n}-\binom{r+n-4}{n}.$$

The proof of Theorem 3.2 is based on Lemma 2.2 about the asymptotic expansion of the solution of the biharmonic equation and the Hardy type inequalities for unbounded domains [6–8]. In case (iv), we need to determine the number of linearly independent solutions of the biharmonic equation (1), the degree of which do not exceed the fixed number.

It is well know that the dimension of the space of all polynomials in $\mathbb{R}^n$ of degree $\leqslant r$ is equal $\binom{r+n}{n}$ [24]. Then the dimension of the space of all biharmonic polynomials in $\mathbb{R}^n$ of degree $\leqslant r$ is equal to

$$\binom{r+n}{n}-\binom{r+n-4}{n},$$

since the biharmonic equation is the vanishing of some polynomial of degree $r-4$ in $\mathbb{R}^n$. If we denote by $k(r,n)$ the number of linearly independent polynomial solutions of equation (1) whose degree do not exceed r and by $l(r,n)$ the number of linearly independent homogeneous polynomials of degree r, that are solutions of equation (1), then

$$k(r,n)=\sum_{s=0}^{r} l(s,n),\quad \text{where}\quad l(s,n)=\binom{s+n-1}{n-1}-\binom{s+n-5}{n-1},\quad s>0.$$

Further, we prove that the mixed Dirichlet–Neumann problem (1), (2) with the condition $D_a(u,\Omega)<\infty$ for $-2r+2-n\leqslant a<-2r+4-n$ has equally $k(r,n)$ linearly independent solutions.

References

[1] F.Brock, An isoperimetric inequality for eigenvalues of the Stekloff problem, *Z. Angew. Math. Mech. (ZAMM)*, **81**(2001), no. 1, 69–71.

[2] R.Farwig, A note on the reflection principle for the biharmonic equation and the Stokes system, *Acta Appl. Math.*, **34**(1994), 41–51.

[3] F.Gazzola, H.-Ch.Grunau, G.Sweers, Polyharmonic Boundary Value Problems: Positivity Preserving and Nonlinear Higher Order Elliptic Equations in Bounded Domains, Lecture Notes Math., vol. 1991, Springer-Verlag, 2010.

[4] D.Gilbarg, N.Trudinger, Elliptic Partial Differential Equations of Second Order. Berlin, Springer-Verlag, 1977.

[5] V.A.Kondratiev, O.A.Oleinik, On the behavior at infinity of solutions of elliptic systems with a finite energy integral, *Arch. Rational Mech. Anal.*, **99**(1987), no. 1, 75–99.

[6] V.A.Kondratiev, O.A.Oleinik, Boundary value problems for the system of elasticity theory in unbounded domains. Korn's inequalities, *Russian Math. Surveys*, **43**(1988), no. 5, 65–119.

[7] V.A.Kondratiev, O.A.Oleinik, Hardy's and Korn's Inequality and their Application, *Rend. Mat. Appl.*, Serie VII, **10**(1990), no. 3, 641–666.

[8] A.A.Kon'kov, On the dimension of the solution space of elliptic systems in unbounded domains, *Russian Acad. Sci. Sbornik Math.*, **80**(1995), no. 2, 411–434.

[9] J.R.Kuttler, V.G.Sigillito, Inequalities for membrane and Stekloff eigenvalues, *J. Math. Anal. Appl.*, **23**(1968), no. 1, 148–160.

[10] O.A.Matevosyan, On solutions of boundary value problems for a system in the theory of elasticity and for the biharmonic equation in a half–space, *Diff. Equations.*, **34**(1998), no. 6, 803–808.

[11] O.A.Matevosyan, The exterior Dirichlet problem for the biharmonic equation: Solutions with bounded Dirichlet integral, *Math. Notes*, **70**(2001), no. 3, 363–377.

[12] O.A.Matevosyan, Solutions of exterior boundary value problems for the elasticity system in weighted spaces, *Sbornik Math.*, **192**(2001), no. 12, 1763–1798.

[13] O.A.Matevosyan, On solutions of mixed boundary-value problems for the elasticity system in unbounded domains, *Izvestiya Math.*, **67**(2003), no. 5, 895–929.
DOI: 10.1070/IM2003v067n05ABEH000451

[14] O.A.Matevosyan, On solutions of the Dirichlet problem for the polyharmonic equation in unbounded domains, *P-Adic Numbers, Ultrametric Analysis, and Appl.*, **7**(2015), no. 1, 74–78. DOI: 10.1134/S2070046615010069

[15] O.A.Matevosyan, Solution of a mixed boundary value problem for the biharmonic equation with finite weighted Dirichlet integral, *Diff. Equations*, **51**(2015), no. 4, 487–501.
DOI: 10.1134/S0012266115040060

[16] O.A.Matevosyan, On solutions of the Neumann problem for the biharmonic equation in unbounded domains, *Math. Notes*, **98**(2015), 990–994. DOI: 10.1134/S0001434615110334

[17] O.A.Matevosyan, On solutions of the mixed Dirichlet–Navier problem for the polyharmonic equation in exterior domains, *Russ. J. Math. Phys.*, **23**(2016), no. 1, 135–138.
DOI: 10.1134/S106192081601012X

[18] O.A.Matevosyan, On solutions of one boundary value problem for the biharmonic equation, *Diff. Equations*, **52**(2016), no. 10, 1379–1383. DOI: 10.1134/S0012266116100153

[19] H.A.Matevossian, On the biharmonic Steklov problem in weighted spaces, *Russ. J. Math. Phys.*, **24**(2017), no. 1, 134–138. DOI: 10.1134/S1061920817010125

[20] H.A.Matevossian, On solutions of the mixed Dirichlet–Steklov problem for the biharmonic equation in exterior domains, *P-Adic Numbers, Ultrametric Analysis, and Appl.*, **9**(2017), no. 2, 151–157. DOI: 10.1134/S2070046617020054

[21] H.A.Matevossian, On the Steklov–type biharmonic problem in unbounded domains, *Russ. J. Math. Phys.*, **25**(2018), no. 2, 271–276. DOI: 10.1134/S1061920818020115

[22] O.A.Matevossian, Mixed Dirichlet–Steklov problem for the biharmonic equation in weighted spaces, *J. Math. Sci. (N. Y.)*, **234**(2018), no 4, 440–454. DOI: 10.1007/s10958-018-4021-8

[23] H.A.Matevossian, On the polyharmonic Neumann problem in weighted spaces, *Complex Variables and Elliptic Equations*, **64**(2019), no. 1, 1–7.
DOI: 10.1080/17476933.2017.1409740

[24] S.G.Mikhlin, Linear Partial Differential Equations, Vyssaya Shkola, Moscow, 1977 (in Russian).

[25] L.E.Payne, Some isoperimetric inequalities for harmonic functions, *SIAM J. Math. Anal.*, **1**(1970), no. 3, 354–359.

[26] S.L.Sobolev, Applications of Functional Analysis in Mathematical Physics, Amer. Math. Soc., Providence, 1991.

[27] W.Stekloff, Sur les problemes fondamentaux de la physique mathematique, *Ann. Sci. de l'E.N.S.*, 3^e serie, **19**(1902), 191–259 et 455–490.

Смешанная бигармоническая задача Дирихле–Неймана во внешних областях

Овик А. Матевосян
Федеральный исследовательский центр «Информатика и управление» РАН
Москва, Российская Федерация
Московский авиационный институт (национальный исследовательский университет)
Москва, Российская Федерация

Аннотация. Изучаются вопросы единственности решения смешанной задачи Дирихле–Неймана для бигармонического уравнения во внешности компактного множества, в предположении, что обобщенное решение этой задачи обладает конечным интегралом Дирихле с весом $|x|^a$. В зависимости от значения параметра a доказаны теоремы единственности (неединственности), и найдены точные формулы для вычисления размерности пространства решений смешанной задачи Дирихле–Неймана.

Ключевые слова: бигармонический оператор, задача Дирихле-Неймана, весовой интеграл Дирихле.

DOI: 10.17516/1997-1397-2020-13-6-763-773
УДК 517.9

Filtration of Liquid in a Non-isothermal Viscous Porous Medium

Alexander A. Papin*
Margarita A. Tokareva†
Rudolf A. Virts‡
Altai State University
Barnaul, Russian Federation

Received 10.08.2020, received in revised form 09.09.2020, accepted 20.10.2020

Abstract. The solvability of the initial-boundary value problem is proved for the system of equations of one-dimensional unsteady fluid motion in a heat-conducting viscous porous medium.

Keywords: Darcy's law, poroelasticity, filtration, solvability, thermal conductivity.

Citation: A.A. Papin, M.A. Tokareva, R.A. Virts, Filtration of Liquid in a Non-isothermal Viscous Porous Medium, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 763–773.
DOI: 10.17516/1997-1397-2020-13-6-763-773.

1. Problem Statement

The urgency of a theoretical study of filtration problems in porous media is associated with their wide application in solving important practical problems: filtration near river dams, reservoirs and other hydraulic structures; movement of magma in the earth's crust, etc. In many practical problems the porosity of the medium is variable, and the medium is deformed. The model of fluid filtration in a viscous non-isothermal porous medium considered in the work is based on the laws of conservation of masses and energy, Darcy's law, as well as rheological relationships for porosity and pressures. The system of equations has the following form [1, 2]:

$$\frac{\partial(1-\phi)\rho_s}{\partial t}+\frac{\partial}{\partial x}((1-\phi)\rho_s v_s)=0,\quad \frac{\partial(\rho_f\phi)}{\partial t}+\frac{\partial}{\partial x}(\rho_f\phi v_f)=0, \tag{1}$$

$$\phi(v_f-v_s)=-\frac{K(\phi)}{\mu}(\frac{\partial p_f}{\partial x}-\rho_f g),\quad \frac{\partial v_s}{\partial x}=-\frac{1}{\xi(\phi,\theta)}p_e, \tag{2}$$

$$\frac{\partial p_{tot}}{\partial x}=-\rho_{tot}g,\quad \rho_{tot}=\phi\rho_f+(1-\phi)\rho_s,\quad p_e=p_{tot}-p_f,\quad p_{tot}=\phi p_f+(1-\phi)p_s, \tag{3}$$

$$(\rho_f c_f\phi+\rho_s c_s(1-\phi))\frac{\partial\theta}{\partial t}+(\rho_f c_f\phi v_f+\rho_s c_s(1-\phi)v_s)\frac{\partial\theta}{\partial x}=\frac{\partial}{\partial x}(\lambda\frac{\partial\theta}{\partial x}), \tag{4}$$

and is solved in the domain $(x,t)\in Q_T=\Omega\times(0,T)$, $\Omega=(0,1)$, under the boundary and initial conditions

*papin@math.asu.ru https://orcid.org/0000-0001-7510-9164
†tma25@mail.ru https://orcid.org/0000-0002-7162-342X
‡virtsrudolf@gmail.com

$$v_s \mid_{x=0,x=1} = v_f \mid_{x=0,x=1} = \frac{\partial \theta}{\partial x} \mid_{x=0,x=1} = 0, \quad \phi \mid_{t=0} = \phi^0(x), \quad \theta \mid_{t=0} = \theta^0(x). \tag{5}$$

This initial-boundary value problem describes the one-dimensional motion of a two-phase medium between impenetrable heat-insulated walls [1, 2]. Here ρ_s, ρ_f, v_s, v_f, are, respectively, the constant real densities and velocities of phases (s is solid porous medium, f is liquid), ϕ is porosity (fraction of pores), p_s and p_f are pressures in solid and liquid phases, p_{tot} is total medium pressure, p_e is effective pressure, ρ_{tot} is two-phase density, θ is absolute temperature, g isdensity of the mass forces, c_s and c_f are heat capacities for at constant volume of phases, $K(\phi)$ is permeability coefficient, μ is dynamic fluid viscosity, $\xi(\phi, \theta)$ is bulk viscosity coefficient, $\lambda(\phi)$ is heat conductivity coefficient (the prescribed functions). The problem is written in Euler coordinates (x, t).

For the permeability coefficient $K(\phi)$, a well-known dependence of the form is used $K(\phi) = K'\phi^n$, where $K' = \text{const} > 0$, $n = 3$ [1]. The bulk viscosity coefficient $\xi(\phi, \theta)$ is usually taken as $\xi(\phi, \theta) = \eta(\theta)/\phi^m$, $m \in [0, 2]$, where $\eta(\theta)$ is the coefficient of dynamic viscosity of the skeleton, which characterizes the relationship between the strain rate tensor and the stress tensor and is determined from the experiment under uniaxial compression [3, 4]. The following dependence is taken as a model one: $\eta(\theta) = \eta_r \exp(Q_r(1-\theta/\theta_r)/R\theta)$, η_r, Q_r, θ_r, R are positive constants (analog of the Arrhenius formula for the dependence of the reaction rate on temperature) [1]. The thermal conductivity coefficient of the medium $\lambda(\phi)$ is taken in the form $\lambda(\phi) = \lambda_f\phi + \lambda_s(1-\phi)$, where λ_f, λ_s are the thermal conductivity of liquid and solid phase (averaged thermal conductivity) [2]. In what follows, the notations are used $k(\phi) = K(\phi)/\mu$, $1/\xi(\phi, \theta) = a_1(\phi)\xi_1(\theta)$, $a_1(\phi) = \phi^m$, $\xi_1(\theta) = 1/\eta(\theta)$.

The local in time solvability of the initial-boundary value problem for the equations (1)–(3) at constant temperature in the case of a compressible fluid was established in the work [5]. A numerical analysis of the initial-boundary value problem for the system (1)–(3) is carried out in [6]: difference schemes are constructed and their convergence is established. In paper [7], the global solvability of the problem (1)– (3) is proved in the case of constant phase densities.

Systems of equations similar in structure were considered in [8–16]. The local solvability of the Cauchy problem in Sobolev spaces was established in [8]. The simplest models of deformation of a poroelastic medium were studied in [9, 10]. Self-similar solutions of the traveling wave type for the equations of magma motion were considered in [11, 12]. The works [14, 15] are devoted to numerical calculations. The problem of substantiating multidimensional models of fluid filtration in poroelastic media is open.

In the notation of function spaces, we follow [15]: $C^{l+\alpha,r+\beta}(Q_T)$ is the Hölder space, where l, r are natural numbers, $(\alpha, \beta) \in (0, 1]$, with the norm $||f||_{C^{l+\alpha,r+\beta}(Q_T)}$.

In this paper, we prove the local classical solvability of the problem (1)–(4) in the case when the bulk viscosity coefficient ξ is a function of porosity and temperature. An example of decidability "in the whole" is given.

Definition. *By a solution of problem (1)–(5) we mean the set of functions $\phi, \phi_t, \theta, v_s, v_f \in C^{2+\alpha,1+\beta}(Q_T)$, $p_f, p_s \in C^{1+\alpha,1+\beta}(Q_T)$, such that $0 < \phi < 1$, $0 < \theta < \infty$. These functions satisfy the equations (1)–(4) and the initial and boundary conditions (5) and regarded as continuous functions in Q_T.*

Theorem 1. *Suppose that the data of problem (1)–(5) satisfies the following conditions:*

1) the functions $k(\phi), a_1(\phi), \lambda(\phi), \xi_1(\theta)$ and their derivatives up to the second order are continuous for $\phi \in (0, 1)$, $\theta \in (0, \infty)$ and satisfy the conditions

$$k_0^{-1}\phi^{q_1}(1-\phi)^{q_2} \leqslant k(\phi) \leqslant k_0\phi^{q_3}(1-\phi)^{q_4},$$

$$k_0^{-1}\phi^{q_5}(1-\phi)^{q_6} \leqslant \lambda(\phi) \leqslant k_0\phi^{q_7}(1-\phi)^{q_8}, \quad \xi_1(\theta) > 0, \quad \theta \in (0,\infty),$$
$$\frac{1}{\xi(\phi)} = a_0(\phi)\phi^{\alpha_1}(1-\phi)^{\alpha_2-1}, \quad 0 < R_1 \leqslant a_0(\phi) \leqslant R_2 < \infty,$$

where $k_0, \alpha_i, R_i, i = 1,2$ are positive constants, $q_1, ..., q_8$ are fixed real numbers.

2) the function g, the initial functions ϕ^0 and θ^0 satisfy the following smoothness conditions:

$$g \in C^{1+\alpha,1+\beta}(\bar{Q}_T), \quad \theta^0, \phi^0 \in C^{2+\alpha}(\bar{\Omega}),$$

and the inequalities

$$0 < m_0 \leqslant \phi^0(x) \leqslant M_0 < 1, \quad 0 < m \leqslant \theta^0(x) \leqslant M < \infty, \quad |g(x,t)| \leqslant g_0 < \infty, \quad x \in \bar{\Omega}, \quad t \in (0,T),$$

where m_0, M_0, m, M, g_0 are given positive constants.

Then problem (1)–(5) has a local solution, i.e., there exists a value of t_0 such that $\phi(x,t), \phi_t(x,t), \theta(x,t) \in C^{2+\alpha,1+\beta}(\bar{Q}_{t_0})$, $(v_s(x,t), v_f(x,t)) \in C^{2+\alpha,\beta}(\bar{Q}_{t_0})$, $(p_f(x,t), p_s(x,t)) \in C^{1+\alpha,\beta}(\bar{Q}_{t_0})$.

Moreover, $0 < \phi(x,t) < 1$, $0 < \theta(x,t) < \infty$ in $\bar{Q}_{t_0}$.

Theorem 2. *Let, in addition to the conditions of Theorem 1, the functions $k(\phi), \xi(\phi,\theta)$ satisfy the conditions*

$$k(\phi) = \frac{K}{\mu}, \quad \xi(\phi,\theta) = \frac{\eta(\theta)}{\phi},$$

where K, μ are positive constants.

Then for all $t \in [0,T]$, $T < \infty$ uniqueness solution of problem (1)–(5) exists, and there are numbers $0 < m_1 < M_1 < 1$, $0 < m_2 < M_2$ such that $m_1 \leqslant \phi(x,t) \leqslant M_1$, $m_2 \leqslant \theta(x,t) \leqslant M_2$, $(x,t) \in Q_T$.

2. Local solvability

Proof of Theorem 1. When proving Theorems 1 and 2, it is convenient to use the Lagrange variables [17]. Suppose that $\bar{x} = \bar{x}(\tau, x, t)$ is a solution of the Cauchy problem

$$\frac{\partial \bar{x}}{\partial \tau} = v_s(\bar{x}, \tau), \quad \bar{x} \mid_{\tau=t} = x.$$

We set $\hat{x} = \bar{x}(0,x,t)$ and take $\hat{x}$ and t for the new variables. Then $\hat{J}(\hat{x},t) = \dfrac{\partial \hat{x}}{\partial x}(x,t) = (1-\phi(\hat{x},t))/(1-\phi^0(\hat{x}))$ is the Jacobian of the transformation. Following [5], we rewrite the system (1)–(4):

$$\frac{\partial}{\partial t}\left(\frac{\phi}{1-\phi}\right) = \frac{\partial}{\partial x}\left(k(\phi)(1-\phi)\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G(\phi)}{\partial t}\right) - k(\phi)g(\rho_{tot}+\rho_f)\right), \tag{6}$$

$$\left((1-\phi)\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right) - g(\rho_{tot}+\rho_f)\right) \mid_{x=0,x=1} = 0, \quad \phi \mid_{t=0} = \phi^0(x), \tag{7}$$

$$\left(c_s\rho_s + c_f\rho_f\frac{\phi}{1-\phi}\right)\frac{\partial \theta}{\partial t} + c_f\rho_f\phi(v_f - v_s)\frac{\partial \theta}{\partial x} = \frac{\partial}{\partial x}\left(\lambda(1-\phi)\frac{\partial \theta}{\partial x}\right), \tag{8}$$

$$\frac{\partial \theta}{\partial x} \mid_{x=0,x=1} = 0, \quad \theta \mid_{t=0} = \theta^0(x), \tag{9}$$

$$\frac{\partial G(\phi)}{\partial t} = \xi_1(\theta)p_e, \quad \frac{dG}{d\phi} = \frac{1}{a_1(\phi)(1-\phi)}. \tag{10}$$

In the system (6)–(10), the basic equations are (6) and (8) for the required functions ϕ and θ.

We substitute in the coefficients of the equation (6) and the boundary condition (7) instead of $\theta(x,t)$ an arbitrary smooth function $\theta_0(x,t) \in C^{2+\alpha_1,1+\beta_1}(Q_T)$, which satisfies the inequalities $0 < m \leqslant \theta^0(x) \leqslant M < \infty$. We retain the previous notation ϕ for solving the arising problem and the latter is called Problem I.

Lemma 1. *Let the data of problem I satisfy the conditions of the theorem. Then problem I has a unique local solution, i.e., there exists a value of t_0 such that*

$$(\phi, \phi_t) \in C^{2+\alpha,1+\beta}(Q_{t_0}),\ \phi \in (0,1).$$

Proof. Suppose that $z = \dfrac{1}{\xi_1(\theta_0)}\dfrac{\partial G}{\partial t}$, we arrive at the following problem for G, z :

$$z = \frac{1}{\xi_1(\theta_0)}\frac{\partial G}{\partial t}, \quad G\,|_{t=0} = G(\phi^0) = G^0(x), \tag{11}$$

$$\frac{z}{d(G,\theta_0)} - \frac{\partial}{\partial x}\left(a(G)\frac{\partial z}{\partial x} - b(G)\right) = 0, \quad \left(a(G)\frac{\partial z}{\partial x} - b(G)\right)|_{x=0,x=1} = 0, \tag{12}$$

where

$$d(G,\theta_0))) = \frac{1-\phi(G)}{a_1(\phi(G))\xi_1(\theta_0)}, \quad a(G) = k(\phi(G))(1-\phi(G)), \quad b(G) = k(\phi(G))g(\rho_{tot}+\rho_f).$$

Since $0 < m_0 \leqslant \phi^0(x) \leqslant M_0 < 1$ and the function $G(\phi)$ is monotone, then $G(m_0) \leqslant G^0(x) \leqslant G(M_0)$. From (11) when the inequality $\max_{(x,t)} |\xi_1(\theta) z(x,t)| \leqslant c_0$ we have that there is a value t_0, such that for all $t \leqslant t_0$ the estimates take place

$$\begin{gathered} G_1(m_0) = G(m_0) - c_0 t_0 \leqslant G(x,t) \leqslant G(M_0) + c_0 t_0 = G_2(M_0), \\ 0 \leqslant G^{-1}(G_1(m_0)) \leqslant \phi(x,t) \leqslant G^{-1}(G_2(M_0)) < 1. \end{gathered} \tag{13}$$

Let $G_0(x,t)$ be a function continuous in x and t, satisfying inequalities (13) and having a continuous derivative $\partial G_0/\partial x$ with respect to x, t. Substituting $G_0(x,t)$ instead of $G(x,t)$ into the coefficients of the equation (12) and the boundary conditions, we arrive at a linear problem for z, in which $a > 0$, $b > 0$ and $d > 0$. The solution to this problem is unique. Existence follows, for example, from Hilbert's theorem [18] for ordinary linear equations of the second order. The t variable plays the role of a parameter. Thus, $(z, z_x, z_{xx}) \in C(Q_{t_0})$. After finding $z(x,t)$, we can find a new value $G(x,t)$ from the equation (11). This value will satisfy the condition (13).

To prove the solvability of problem I, we use the method of successive approximations. Let $z^i(x,t)$ and $G^i(x,t)$ be a solution to the problem

$$\begin{gathered} \frac{\partial G^{i+1}}{\partial t} = \xi_1(\theta_0) z^{i+1}, \quad G^{i+1}(x,0) = G^0(x), \\ \frac{z^{i+1}}{d(G^i)} - \frac{\partial}{\partial x}\left(a(G^i)\frac{\partial z^{i+1}}{\partial x} - b(G^i)\right) = 0, \\ \left(a(G^i)\frac{\partial z^{i+1}}{\partial x} - b(G^i)\right)|_{x=0,x=1} = 0, \end{gathered}$$

where $i = 0, 1, 2, \ldots$. Substituting $G^0(x)$ into the equation for z at the first step, we find $z^1(x,t)$. After that, from the equation for G we find $G^1(x,t)$, etc. For each i there is a unique solution

$z^i(x,t)$ and $G^i(x,t)$, satisfying (13). It is checked in a standard way that for a small value of t_0 the solutions $z^i(x,t)$, $G^i(x,t)$ and their derivatives up to the second order inclusive are bounded uniformly in i.

We put $y^{i+1} = z^{i+1} - z^i$, $\omega^{i+1} = G^{i+1} - G^i$. We have

$$\frac{\partial \omega^{i+1}}{\partial t} = \xi_1(\theta_0) y^{i+1}, \quad \omega^{i+1} \mid_{t=0} = 0,$$

$$\frac{y^{i+1}}{d(G^i)} + A_1 \omega^i - \frac{\partial}{\partial x}\left(a y_x^{i+1} + A_2 \omega^i\right) = 0,$$

$$(a y_x^{i+1} + A_2 \omega^i)|_{x=0,x=1} = 0,$$

where the coefficients A_1, A_2 are easily recovered and are limited. We have from this system the following inequalities

$$\int_0^1 (|y^{i+1}|^2 + |y_x^{i+1}|^2) dx \leqslant c_1 \int_0^1 |\omega^i|^2 dx \leqslant c_1 \max_x |\omega^i|^2,$$

$$\max_x |\omega^{i+1}| \leqslant c_1 \int_0^t \max_x |y^{i+1}| d\tau,$$

where the constant c_1 does not depend on i. Taking into account the last inequality for the function $v^i(t) = \max_x |y^i(x,t)|^2$ we get $v^{i+1}(t) \leqslant c_2 \int\limits_0^t v^i(\tau) d\tau$ and therefore [19], $v^i(t) \leqslant (c_2 T)^i v^0 / i! \to 0$ for $i \to \infty$. After that it is easy to establish that the sequences z^i, G^i are fundamental in $C(Q_{t_0})$ and have limits $z(x,t) \in C(Q_{t_0})$ and $G(x,t) \in C(Q_{t_0})$. The sequences z_x^i, z_{xx}^i, G_t^i are also fundamental. Passing to the limit as $i \to \infty$, we obtain that the limit functions satisfy the problem (11), (12). The uniqueness of the solution is proved similarly to [7]. Increasing the smoothness of the initial data to those specified in the conditions of Theorem 1 allows us to obtain that $\phi(x,t), \phi_t(x,t) \in C^{2+\alpha,1+\beta}(\bar{Q}_{t_0})$.

Lemma 1 is proved. □

Substituting $\theta_0(x,t)$ and the solution to Problem I into the coefficients of equation (8), we arrive at a linear problem for $\theta(x,t)$ of the form

$$Q \frac{\partial \theta}{\partial t} + V \frac{\partial \theta}{\partial x} = \frac{\partial}{\partial x}\left(\lambda(1-\phi) \frac{\partial \theta}{\partial x}\right),$$

$$\frac{\partial \theta}{\partial x} \mid_{x=0,x=1} = 0, \quad \theta \mid_{t=0} = \theta^0(x),$$

where

$$Q = \rho_s c_s + \rho_f c_f \frac{\phi}{1-\phi}, \quad V = c_f \rho_f \phi (v_f - v_s) = \rho_f c_f k(\phi) \left((1-\phi)\frac{\partial z}{\partial x} + g(\rho_{tot} + \rho_f)\right).$$

The unique solvability of this problem in Holder classes follows from [19], and the solution satisfies the estimate

$$0 < \underline{\theta} = \min_x \theta^0(x) \leqslant \theta(x,t) \leqslant \max_x \theta^0(x) = \bar{\theta} < \infty.$$

After these remarks, the local solvability of the problem (6)–(9) can easily be obtained using the Schauder theorem according to the scheme used in [7].

After finding ϕ, θ, the remaining functions from the system (1)–(4) can be defined as follows. We find the phase velocities from (1)

$$v_f(x,t) = -\frac{1}{\phi}\int_0^x \frac{\partial \phi}{\partial t} d\xi \in C^{2+\alpha,\beta}(Q_{t_0}),$$

$$v_s(x,t) = -\frac{1}{1-\phi}\int_0^x \frac{\partial (1-\phi)}{\partial t} d\xi \in C^{2+\alpha,\beta}(Q_{t_0}).$$

From (3) we find $p_{tot}(x,t) = p^0(t) - \int\limits_0^x \rho_{tot} g d\xi \in C^{3+\alpha,1+\beta}(Q_{t_0})$.

From (2) we have $p_e(x,t) = -\dfrac{\partial v_s}{\partial x}\xi(\phi,\theta) \in C^{1+\alpha,\beta}(Q_{t_0})$, then

$$p_f(x,t) = p_{tot} - p_e \in C^{1+\alpha,\beta}(Q_{t_0}), \quad p_s(x,t) = \frac{p_{tot}}{1-\phi} - \frac{\phi}{1-\phi} p_f \in C^{1+\alpha,\beta}(Q_{t_0}).$$

Theorem 1 is proved. □

3. Global solvability

Proof of Theorem 2. By Theorem 1, we will assume that on the interval $[0, t_0]$ there exists a solution to the problem (1)–(5), and $0 < \phi(x,t) < 1$, $0 < \theta(x,t) < \infty$, $x \in \Omega$, $t \in [0, t_0]$. After obtaining the necessary a priori estimates that do not depend on the value of t_0, the local solution can be continued to the entire segment $[0, T]$.

Lemma 2. *Under the conditions of Theorem 2, for all $t \in [0,T]$ the following relations hold:*

$$\int_0^1 s(x,t)dx = \int_0^1 s^0(x)dx, \quad s = \frac{\phi}{1-\phi}, \quad s^0 = s(x,0), \tag{14}$$

$$0 < \underline{\theta} \equiv \min_{x\in[0,1]} \theta^0(x) \leqslant \theta(x,t) \leqslant \max_{x\in[0,1]} \theta^0(x) \equiv \overline{\theta} < \infty, \tag{15}$$

$$\int_0^1 \frac{1}{\xi_1(\theta)} \frac{a_1}{1-\phi} \left(\frac{\partial G}{\partial t}\right)^2 dx + \frac{1}{2}\int_0^1 k(\phi)(1-\phi)\left|\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)\right|^2 dx \leqslant$$
$$\leqslant \frac{1}{2}\int_0^1 \frac{k(\phi)}{1-\phi} g^2 (\rho_{tot} + \rho_f)^2 dx \leqslant N. \tag{16}$$

Hereinafter, N denotes a constant that depends only on the data of the problem (1)–(5) and does not depend on t_0.

Proof. Let us integrate the equation (6) over x from 0 to 1 and take into account the boundary condition (7). After integration over time from 0 to the current value of t, we arrive at the equality (14).

The equation (8) is written in a divergent form:

$$\frac{\partial}{\partial t}\left(\theta(c_s\rho_s + c_f\rho_f\frac{\phi}{1-\phi})\right) + \frac{\partial}{\partial x}\left(\theta c_f\rho_f\phi(v_f - v_s) - \lambda(1-\phi)\frac{\partial\theta}{\partial x}\right) =$$

$$= \theta\left[\frac{\partial}{\partial t}\left(c_s\rho_s + c_f\rho_f\frac{\phi}{1-\phi}\right) + \frac{\partial}{\partial x}(c_f\rho_f\phi(v_f - v_s))\right]. \tag{17}$$

The right-hand side of this equality is equal to zero, since the second equation from (1) in Lagrange variables becomes [5]

$$\frac{\partial}{\partial t}\left(\frac{\phi}{1-\phi}\right)+\frac{\partial}{\partial x}\left(\phi(v_f-v_s)\right)=0.$$

In particular, from (17) we have

$$\int_0^1\left(c_f\rho_f\frac{\phi}{1-\phi}+c_s\rho_s\right)\theta dx=\int_0^1\left(c_f\rho_f\frac{\phi^0}{1-\phi^0}+c_s\rho_s\right)\theta^0 dx,$$

and therefore $\theta(x,t)\in L_1[0,1]$ for all $t\in[0,T]$.

Let the smooth function $\kappa(\theta)$ satisfy the condition $\kappa''(\theta)=d^2\kappa/d\theta^2\geqslant 0$. Multiplying the equation (8) by $\kappa'(\theta)=d\kappa/d\theta$, and following the equality (17) we reduce the resulting equality to the form

$$\frac{\partial}{\partial t}\left(\left(c_s\rho_s+c_f\rho_f\frac{\phi}{1-\phi}\right)\kappa(\theta)\right)+\frac{\partial}{\partial x}(c_f\rho_f\phi(v_f-v_s)\kappa(\theta))=$$
$$=\frac{\partial}{\partial x}\left(\lambda(1-\phi)\frac{\partial\kappa(\theta)}{\partial x}\right)-\kappa''(\theta)\left(\frac{\partial\theta}{\partial x}\right)^2\lambda(1-\phi). \tag{18}$$

In the case $\kappa(\theta)=\theta^p$, $p>1$, from (18) we deduce

$$\int_0^1\theta^p(x,t)dx\leqslant\max_{x\in[0,1]}\left(\frac{c_f\rho_f}{c_s\rho_s}\frac{\phi^0(x)}{1-\phi^0(x)}+1\right)\int_0^1|\theta^0(x)|^p dx.$$

Whence, in the standard way, we get that $\theta(x,t)\leqslant\max_{x\in[0,1]}\theta^0(x)$ for all $t\in[0,T]$, $x\in[0,1]$. Put $\theta_1=1/\theta$ and the equation (6) can be represented as‘

$$\left(c_s\rho_s+c_f\rho_f\frac{\phi}{1-\phi}\right)\frac{\partial\theta_1}{\partial t}+c_f\rho_f(v_f-v_s)\frac{\partial\theta_1}{\partial x}=\frac{\partial}{\partial x}\left(\lambda(1-\phi)\frac{\partial\theta_1}{\partial x}\right)-2\lambda(1-\phi)\left(\frac{\partial\theta_1}{\partial x}\right)^2\theta.$$

Multiplying (8) by $\kappa_1'(\theta_1)=d\kappa_1/d\theta_1$, $\kappa_1=\theta_1^p$, and integrating over x, we arrive at a relation of the form (14) for $\theta_1(x,t)$. Therefore $\theta(x,t)\geqslant\min_{x\in[0,1]}\theta^0(x)$ for all $t\in[0,T], x\in[0,1]$.

Multiplying the equation (6) by $\dfrac{1}{\xi_1(\theta)}\dfrac{\partial G}{\partial t}$ and integrating over x we arrive at the relation

$$\int_0^1\frac{1}{\xi_1(\theta)}\frac{a_1(\phi)}{1-\phi}\left(\frac{\partial G}{\partial t}\right)^2dx+\int_0^1k(\phi)(1-\phi)|\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)|dx=$$
$$=\int_0^1k(\phi)g(\rho_{tot}+\rho_f)\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)dx\leqslant$$
$$\leqslant\frac{1}{2}\int_0^1k(\phi)(1-\phi)\left|\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)\right|^2dx+\frac{1}{2}\int_0^1\frac{k(\phi)}{1-\phi}g^2(\rho_{tot}+\rho_f)^2dx.$$

The last term on the right-hand side is bounded uniformly in t_0, since $\phi<1$ and, therefore, $\rho_{tot}\leqslant\max(\rho_f,\rho_s)$. Finally, due to (14) we have

$$\int_0^1\frac{dx}{1-\phi}=1+\int_0^1s^0(x)dx.$$

Lemma 2 is proved. □

Lemma 3. *Under the conditions of Theorem 2, for all $t \in [0,T]$, $x \in [0,1]$ the estimate takes place*

$$0 < m \leqslant \phi(x,t) \leqslant M < 1. \tag{19}$$

Proof. From the inequality (16) by the conditions of Theorem 2 it follows

$$\int_0^1 \left| \frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)\right| dx \leqslant \left(\int_0^1 \frac{dx}{1-\phi}\right)^{1/2} \left(\int_0^1 (1-\phi)\left|\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)\right|^2 dx\right)^{1/2}.$$

From (6) it also follows that

$$\int_0^1 \frac{a_1}{1-\phi}\frac{\partial G}{\partial t}dx = 0,$$

and, therefore, there is a point $x_0(t)$ at which $\dfrac{\partial G}{\partial t}(x_0(t),t) = 0$. Therefore

$$\min_{x\in(0,1)}\left|\frac{1}{\xi_1(\theta)}\right|\left|\frac{\partial G}{\partial t}\right| \leqslant \left|\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right| \leqslant \int_0^1 \left|\frac{\partial}{\partial x}\left(\frac{1}{\xi_1(\theta)}\frac{\partial G}{\partial t}\right)\right| dx \leqslant N.$$

Taking into account (15) and the conditions of Theorem 2, from the last inequality we have

$$|lns(x,t)| \leqslant |G(x,t)| \leqslant |G^0(x)| + N_1 T \leqslant N_2.$$

Then we arrive at (19) with $m = (1+e^{N_2})^{-1}$, $M = (1+e^{-N_2})^{-1}$.

Let $z = \dfrac{1}{\xi_1(\theta)}\dfrac{\partial G}{\partial t}$. The problem (6), (7) takes the form

$$\frac{a_1(\phi)\xi_1(\theta)z}{(1-\phi)} = \frac{\partial}{\partial x}\left(k(\phi)(1-\phi)\frac{\partial z}{\partial x} - k(\phi)g(\rho_{tot}+\rho_f)\right),$$

$$\left(k(\phi)(1-\phi)\frac{\partial z}{\partial x} - k(\phi)g(\rho_{tot}+\rho_f)\right)|_{x=0,x=1} = 0.$$

By Lemmas 2 and 3, we have

$$\int_0^t\int_0^1 \theta_x^2 dx d\tau + \int_0^1 (z^2 + z_x^2 + \theta_x^2)dx \leqslant N_3,$$

where N_3 is a positive constant depending on the initial data, parameters and problem constants, but does not depend on t_0.

Using the representation

$$G(\phi) = \int_0^t \xi_1(\theta) z d\tau + G(\phi^0),$$

we get

$$G'(\phi)\phi_x = \int_0^t (z_x\xi_1(\theta) + z\xi_1'\theta_x)d\tau + G_x(\phi^0).$$

Therefore

$$\int_0^1 \phi_x^2 dx \leqslant N_4.$$

The equation for function $z(x,t)$ takes form

$$a_0(\phi,\theta)z = a_1(\phi)z_{xx} + a_1'(\phi)\phi_x z_x + a_2'(\phi)\phi_x.$$

The coefficients $a_0(\phi,\theta) > 0,\ a_1(\phi) > 0,\ a_2(\phi)$ are limited and easy to calculate.

We have

$$\int_0^1 z_{xx}^2 dx \leqslant C_1 \left(\int_0^1 (z^2 + \phi_x^2) dx + \int_0^1 |z_{xx} z_x \phi_x| dx \right),$$

where

$$I_1 = \int_0^1 |z_{xx}||z_x\phi_x| dx \leqslant \max|z_x| \left(\int_0^1 z_{xx}^2 dx \right)^{1/2} \left(\int_0^1 \phi_x^2 dx \right)^{1/2} \leqslant$$

$$\leqslant C_1 \left(\left(\int_0^1 z_{xx}^2 dx \right)^{1/2} \left(\int_0^1 \phi_x dx \right)^{1/2} + \left(\int_0^1 z_{xx}^2 dx \right)^{3/4} \left(\int_0^1 \phi_x dx \right)^{1/2} \right).$$

The constant C_1 is not depend on t_0.

Therefore

$$\max_x |z_x| + \int_0^1 z_{xx}^2 dx \leqslant N_4.$$

The equation for the function $\theta(x,t)$ has the form

$$\theta_t + a_3(\phi, z_x)\theta_x = a_4(\phi)\theta_{xx} + a_5(\phi)\phi_x\theta_x,$$

where the coefficients $a_4(\phi) > 0$, $a_3(\phi, z_x)$, $a_5(\phi)$ are limited and easy to calculate.

Since

$$\int_0^1 |\theta_x \theta_{xx} \phi_x| dx \leqslant \max_x |\theta_x| \left(\int_0^1 \theta_{xx}^2 dx \right)^{1/2} \left(\int_0^1 \phi_x^2 dx \right)^{1/2} \leqslant$$

$$\leqslant c \left(\int_0^1 \theta_{xx}^2 dx \right)^{3/4} \left(\int_0^1 \phi_x^2 dx \right)^{1/2} \left(\int_0^1 \theta_x^2 dx \right)^{1/4},$$

then from the equation for θ we have

$$\int_0^1 \theta_x^2 dx + \int_0^t \int_0^1 (\theta_t^2 + \theta_{xx}^2) dx d\tau \leqslant N_5.$$

To complete the proof of Theorem 2, it is necessary to obtain the Holder continuity in x, t of the functions ϕ_x and z_x included in the coefficients of the equations for z and θ. From the embedding $z_{xx} \in L_2[0,1]$ and the representation for ϕ we have $\phi_{xx} \in L_2[0,1]$. Then for $w = \theta_x$ we get

$$\int_0^1 (\theta_t^2 + w_x^2) dx + \int_0^t \int_0^1 (w_t^2 + w_{xx}^2) dx d\tau \leqslant N_6.$$

After that we deduce that $|\phi_{xt}| \leqslant N_7$. Finally, following [7] for the function $\sigma = z_t$ we get $\sigma_x \in L_2[0,1]$.

Theorem 2 is proved. □

Conclusion

The local solvability in the Holder classes of the initial-boundary value problem of one-dimensional fluid motion in a nonisothermal viscous porous medium is proved. An example of decidability is given at any finite time interval.

The work was carried out in accordance with the State Assignment of the Russian Ministry of Science and Higher Education entitled 'Modern methods of hydrodynamics for environmental management, industrial systems and polar mechanics' (Govt. contract code: FZMW-2020-0008, 24 January 2020).

References

[1] J.A.D.Connoly, Y.Y.Podladchikov, Compaction-driven fluid flow in viscoelastic rock, *Geodinamica Acta*, **11**(1998), no 2-3, 55–84.

[2] A.Fowler, Mathematical Geoscience, Springer-Verlag London Limited, 2011.

[3] J.A.D.Connolly, Y.Y.Podladchikov, Temperature-dependent viscoelastic compaction and compartmentalization in sedimentary basins, *Tectonophysics*, **324**(2000), no. 3, 137–168.

[4] R.E.Grimm, S.C.Solomon, Viscous relaxation of impact crater relief on Venus: Constraints on crustal thickness and thermal gradient, *Journal of Geophysical Research: Solid Earth*, **93**(1988), no. B10, 11911–11929.

[5] A.A.Papin, M.A.Tokareva, On Local solvability of the system of the equation of onedimensional motion of magma, *Journal of Siberian Federal University. Mathematics and Physics*, **10**(2017) no. 3, 385–395. DOI: 10.17516/1997-1397-2017-10-3-385-395

[6] M.N. Koleva, L.G.Vulkov, Numerical analysis of one dimensional motion of magma without mass forces, *Journal of Computational and Applied Mathematics*, **366**(2020), 112338. DOI: 10.1016/j.cam.2019.07.003

[7] M.A.Tokareva, A.A.Papin, Global solvability of a system of equations of one-dimensional motion of a viscous fluid in a deformable viscous porous medium, *Journal of Applied and Industrial Mathematics*, **13**(2019), no. 2, 350–362. DOI: 10.33048/sibjim.2019.22.208

[8] M.Simpson, M.Spiegelman, M.I.Weinstein, Degenerate dispersive equations arising in the stady of magma dynamics, *Nonlinearity*, **20**(2007), no. 1, 21–49.

[9] V.Y.Rudyak, O.B.Bocharov, A.V.Seryakov, Hierarchical sequence of models and deformation peculiarities of porous media saturated with fluids, Proceedings of the XLI Summer School-Conference Advanced Problems in Mechanics (APM-2013), 2013, 184–191.

[10] O.B.Bocharov, V.Y.Rudyak, A.V.Seryakov, Simplest deformation models of a fluid-saturated poroelastic medium, *Journal of Mining Science*, **50**(2014), 235–248.

[11] A.M.Abourabia, K.M.Hassan, A.M.Morad, Analytical solutions of the magma equations for molten rocks in a granular matrix, *Chaos, Solitons and Fractals*, **42**(2009), no. 2, 1170–1180.

[12] Y.Geng, L.Zhang, Bifurcations of traveling wave solutions for the magma equation, *Applied Mathematics and computation*, **217**(2010), no. 4, 1741–1748.

[13] I.G.Akhmerova, A.A.Papin, Solvability of the boundary-value problem for equations of one-dimensional motion of a two-phase mixture, *Mathematical Notes*, **96**(2014), no. 2, 166–179.

[14] C.Morency et al., A numerical model for coupled fluid flow and matrix deformation with applications to disequilibrium compaction and delta stability, *Journal of Geophysical Research: Solid Earth*, **112**(2007), no. B10.

[15] A.S.Saad, B.Saad, M.Saad, Numerical study of compositional compressible degenerate twophase flow in saturated-unsaturated heterogeneous porous media, *Comput. Math. Appl.*, **71**(2016), no. 2, 565–584.

[16] S.N.Antontsev, A.V.Kazhikhov, V.N.Monakhov, Boundary value problems of the mechanics of inhomogeneous fluids, Nauka. Sib. branch, 1983 (in Russian).

[17] A.A.Papin, I.G.Akhmerova, Solvability of the system of equations of one-dimensional motion of a heat-conducting two-phase mixture, *Mathematical Notes*, **87**(2010), no. 2, 230–243.

[18] P.I.Lizorkin, A course of differential and integral equations with additional chapters of analysis, Nauka, 1981 (in Russian).

[19] O.A.Ladyzhenskaya, V.A.Solonnikov, and N.N.Ural'tseva, Linear and Quasilinear Equations of Parabolic Type, Moscow, Nauka, 1967 (in Russian).

Фильтрация жидкости в неизотермической вязкой пористой среде

Александр А. Папин
Маргарита А. Токарева
Рудольф А. Вирц
Алтайский государственный университет
Барнаул, Российская Федерация

Аннотация. Для системы уравнений одномерного нестационарного движения жидкости в теплопроводной вязкой пористой среде доказана разрешимость начально-краевой задачи.

Ключевые слова: закон Дарси, пороупругость, фильтрация, разрешимость, теплопроводность.

DOI: 10.17516/1997-1397-2020-13-6-774-780
УДК 512.54

Patterns of Magnetohydrodynamic Flow in the Bent Channel

Alexander V. Proskurin*
Altai State Technical University
Barnaul, Russian Federation
Anatoly M. Sagalakov†
Altai State University
Barnaul, Russian Federation

Received 26.07.2020, received in revised form 05.08.2020, accepted 20.09.2020

Abstract. The article considers the flow patterns of an electrically-conductive fluid in a 90 degree bend. The magnetic field is directed parallel to the outlet branch of the bend. Magnetohydrodynamic equations in terms of the small magnetic Reynolds numbers approach and the spectral-element method were used. The flow patterns were studied at different values of the Reynolds and the Hartmann numbers, and with regard to different values of the bent radus. A reverse flow was found in the outlet branch of the channel.

Keywords: magnetohydrodynamics, channel flows, spectral-element method.

Citation: A.V. Proskurin, A.M. Sagalakov, Patterns of Magnetohydrodynamic Flow in the Bent Channel, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 774–780. DOI: 10.17516/1997-1397-2020-13-6-774-780.

Introduction

The phenomenon of a magnetic field interaction with fluid is observed in nature, and is widely used in industry. Plasma in some cases can be considered as a viscous electrically-conducting fluid. It is intended to use liquid metals for cooling advanced nuclear and thermonuclear reactors, and in large batteries that are designed to buffer energy from wind- and solar-power plants. For the design of such devices, it is important to understand the interaction mechanism of the folded flow of the electrically conducting fluid with the magnetic field. In the cases of a jet and a single vortex influenced by a transverse magnetic field, this issue was considered in [1, 2]. In these papers it was found that a vortex in a uniform transverse magnetic field can generate secondary vortices that rotate in the reverse direction. In [3], it is described that similar phenomena can be observed in a bent channel in a vertical magnetic field in that a reverse flow was observed in the inlet branch. In this paper, we study conditions for the origin of the reverse flow in the bend in the presence of a horizontal magnetic field.

1. Equations and numeric method

Consider the flow in a bent channel as shown in Fig. 1. The length of the input and output branches are indicated as L_1 and L_2 respectively. The channel width is $2d$, and the bend radius is R. The flow of the electrically-conducting viscous fluid occurs under a constant pressure gradient between the "inflow" and "outflow". The state Hartmann flow with the maximal velocity V_0

*k210@list.ru https://orcid.org/0000-0002-4485-2120
†amsagalakov@mail.ru

forms in the inlet branch. The Reynolds number is

$$Re = \frac{V_0 d}{\nu}, \tag{1}$$

where ν is the viscosity. The Hatrmann number is

$$Ha = dB_0\sqrt{\frac{\sigma}{\rho\nu}}, \tag{2}$$

where B_0 is the magnetic field, σ is the electrical conductivity, and ρ is the density of the fluid.

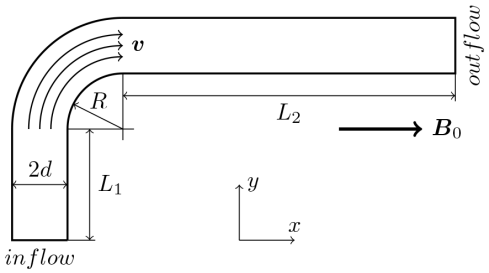


Fig. 1. The bent channel

The problem is considered under the assumption that the magnetic field generated by the movement of the fluid does not affect the flow. This small magnetic Reynolds number approach is suitable for most engineering applications [4]. It is now possible to write the equations in the following form:

$$\begin{aligned} \frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\nabla)\,\boldsymbol{v} &= -\frac{1}{\rho}\nabla p + \nu\Delta\boldsymbol{v} + \boldsymbol{F}(\boldsymbol{v},\boldsymbol{B}_0), \\ \nabla\cdot\boldsymbol{v} &= 0, \end{aligned} \tag{3}$$

where $\boldsymbol{v}$ is the fluid velocity, p is the pressure, and $\boldsymbol{F}$ is the magnetic force.

Omh's law is

$$\boldsymbol{j} = \sigma\left(-\nabla\varphi + \boldsymbol{v}\times\boldsymbol{B}_0\right), \tag{4}$$

where $\boldsymbol{j}$ is the electric current density, φ is the electric potential. A condition $\nabla\cdot\boldsymbol{j} = 0$ for the electric current leads to

$$\Delta\varphi = \nabla\cdot(\boldsymbol{v}\times\boldsymbol{B}_0). \tag{5}$$

Using Reynolds (1) and Hartmann (2) numbers , equations (3) can be written in a non-dimensional form

$$\begin{aligned} \frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v}\cdot\nabla)\,\boldsymbol{v} &= -\nabla p + \frac{1}{Re}\Delta\boldsymbol{v} + \frac{Ha^2}{Re}\left(-\nabla\varphi + \boldsymbol{v}\times\boldsymbol{B}_0\right)\times\boldsymbol{B}_0, \\ \nabla\cdot\boldsymbol{v} &= 0, \\ \Delta\varphi &= \nabla\cdot(\boldsymbol{v}\times\boldsymbol{B}_0). \end{aligned} \tag{6}$$

As the flow is two-dimensional, the flow fields do not depend upon the coordinate z, and $v_z = 0$. This approximation leads to the simplified form $\Delta\varphi = 0$ of the electric potential equation, and hence $\varphi \equiv 0$. Consequently, the magnetic forces will take the form $\boldsymbol{F}(\boldsymbol{v},\boldsymbol{B}_0) = \dfrac{Ha^2}{Re}(\boldsymbol{v}\times\boldsymbol{B}_0)\times\boldsymbol{B}_0$ and the electric potential is excluded from the equations (6).

Boundary conditions $\boldsymbol{v} = 0$ are set on the channel walls. At the inflow, the Hartmann flow is established

$$v_y(x) = \frac{\cosh(Ha) - \cosh(Ha \cdot x)}{\cosh(Ha) - 1}. \tag{7}$$

At the outflow, the velocity satisfies the condition

$$\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{n}} = 0. \tag{8}$$

The spectral-element method and the computer program described earlier in [5] were used for calculations. The mesh is shown in Fig. 2. The state flow was determined by integrating equations over time until constant values of at least eight digits were established at test points. The Fig. 3 shows some of these points, marked with the letters A, B, C. The Tab. 1 contains the velocity values at these points when increasing the approximation order p from 5 to 12. The convergence has been achieved to at least five significant digits. A similar analysis with different grids was performed in [3].

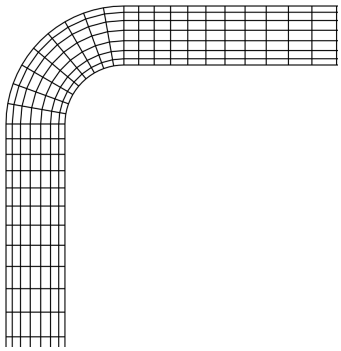


Fig. 2. The mesh

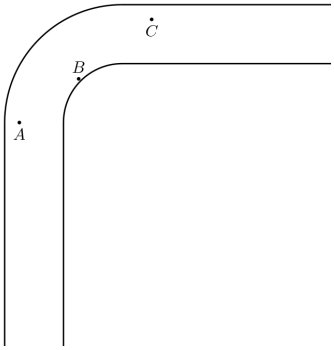


Fig. 3. The test points

Table 1. Convergence at the test points

p	A, v_y	B, v_y	C, v_x	C, v_y
5	0.98470321	0.8643492	–0.03093629	–0.00133713
7	0.98465574	0.8639382	–0.03093710	–0.00131847
10	0.98465842	0.8638807	–0.03093676	–0.00132150
12	0.98465844	0.8638864	–0.03093673	–0.00132158

2. Results and discussion

Primarily, the flow at small Reynolds numbers was considered. Fig. 4 shows streamlines in the outlet branch at $Re = 0.1$ and $Ha = 10$ (a), $Ha = 35$ (b), $Ha = 100$ (c), and $Ha = 300$ (d). At $Ha = 10$, the streamlines are parallel. At $Ha = 35$, a small vortex is observed near the outer wall after the bend. At $Ha = 100$, a reverse flow is observed near the outer wall of the outlet branch. When the Hartmann number increases to $Ha = 300$, the reverse flow area shifts to the center and a return jet is formed near the channel axis. The corresponding velocity profiles in the middle of the length of the outlet branch are shown in Figs. 5 and 6. Fig. 6 shows the velocity at a large scale by a dashed line. The magnitude of the reverse flow has a perceptible value by comparison to the velocity scale V_0.

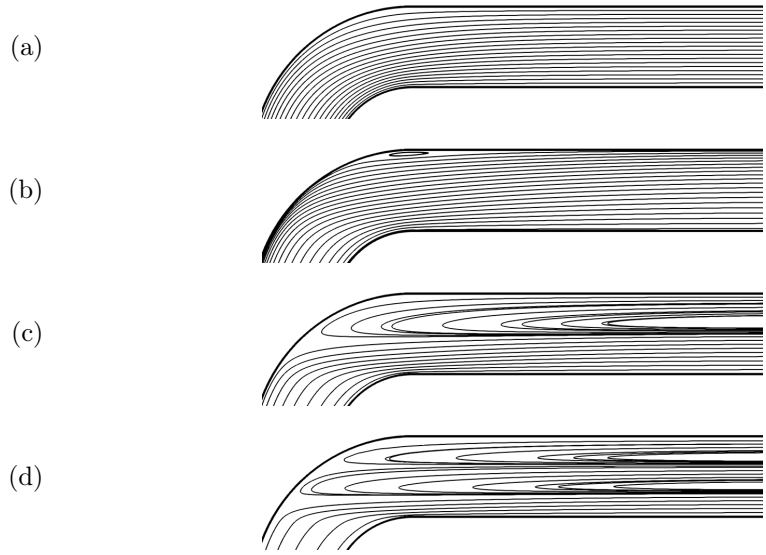


Fig. 4. Streamlines at $Re = 0.1$, $R = 2$: $Ha = 10$ (a), $Ha = 35$ (b), $Ha = 100$ (c), $Ha = 300$ (d)

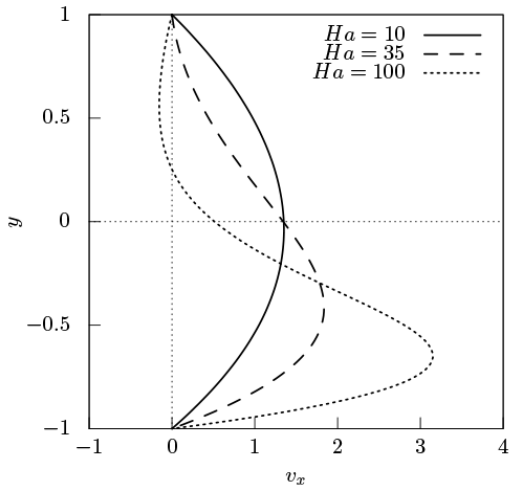


Fig. 5. Velocity v_x in the outlet branch at $R = 2$, $Re = 0.1$: $Ha = 10$, $Ha = 35$, $Ha = 100$.

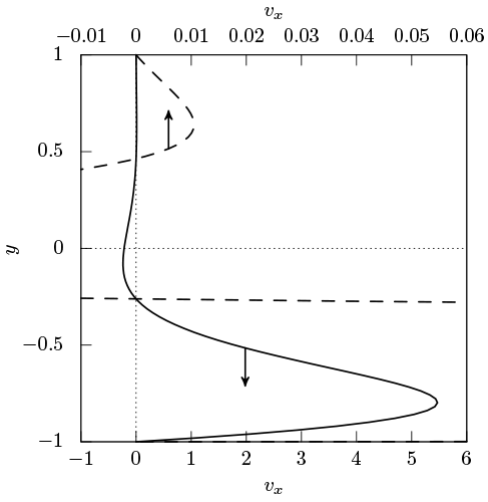


Fig. 6. Velocity v_x in the outlet branch at $R = 2$, $Re = 0.1$, $Ha = 300$. The dashed line shows the plot at the large scale.

Fig. 5 shows that for $Ha = 10$ the velocity profile is symmetrical. When the magnetic field increases, the velocity maximum drifts to the inner wall. To move to the upper part of the outlet branch, the fluid would have to flow across the magnetic field. With regard to this direction, the magnetic force supresses the movement of the fluid. In the inlet branch, this magnetic braking is compensated for by the pressure gradient, but there are no forces that would cause a vertical movement in the outlet pipe. Such forces exist only in the bend, where the velocity distribution is formed due to the action of the inertia forces, the magnetic forces, and the gradient of the pressure field. In Fig. 4 (c), one can see that the streamlines from the input branch take up only

the lower half of the outlet branch. In the upper half, the fluid is dragged by viscous forces, and the flowing forward volume is compensated for by the reverse flow from the outlet of the channel.

The mechanics of motion in a bent channel is similar to that of a free vortex in the transverse field, for which analytical solutions and estimates were obtained in [2]. The origin of the pair of reverse vortices was described in [1, 2] and is analogous to the origin of the reverse flow.

Fig. 7 shows the dependencies of the critical Hartmann number Ha_* from the Reynolds number. The critical Hartmann number is the number at which the reverse flow occurs, initially in the form of a small vortex as shown in Fig. 4 (b). The bend radii were considered equal to $R = 1$, 2 and 3. For the $Re \to 0$, the dependencies $Ha_*(Re)$ have horizontal asymptotes, that is, the occurrence of the reverse flow does not depend on the Reynolds number. At $10 < Re < 100$, these dependencies have minima. For $Re > 100$, the curves $Ha_*(Re)$ increases.

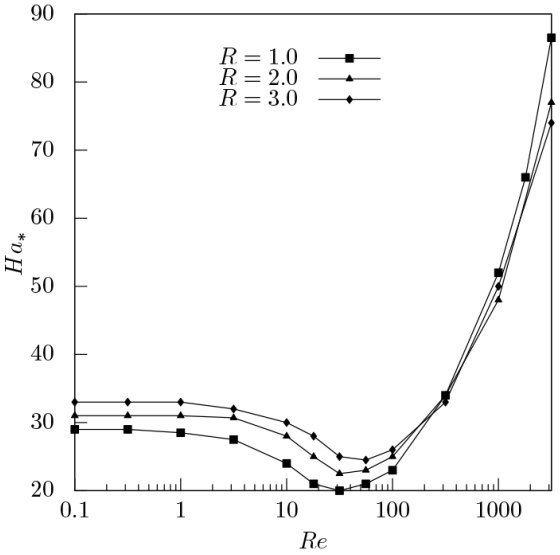


Fig. 7. Reverse flow diagram for $R = 1$, 2, 3

Fig. 8 shows the streamlines of the reverse flow at $Re = 1000$: $Ha = 10$ (a), $Ha = 65$ (b), $Ha = 100$ (c). At $Ha = 10$, a vortex is observed near the inner wall of the channel straightway after the bend. At $Ha = 65$, a vortex exists at the outer wall in the outlet branch. When $Ha = 100$, a reverse flow is observed. At the Reynolds number $Re = 1000$ it was not possible to obtain a flow without vortices, as shown in Fig. 4 (a): at the smallest Hartmann numbers the flow has the form as in Fig. 8 (a). Also, due to instability, it was not possible to obtain a state reverse jet at $Ha = 300$.

Conclusion

This paper describes the flow of a viscous electrically-conducting fluid in a bent channel. The magnetic field is directed parallel to the outlet branch. The occurrence of the reverse flow in the form of the near-wall flow and the near-axial jet is presented, including the data for several bend radii. The obtained results are interesting with regard to the design of magnetohydrodynamic devices such as liquid metal blankets for thermonuclear reactors, given that they have a large number of bent channels. The suppression of rotational motion by a magnetic field should have a strong effect on their hydraulic characteristics. At the same time, the complete mathematical modeling of flows in engineering devices is currently a very expensive task. A solution of model

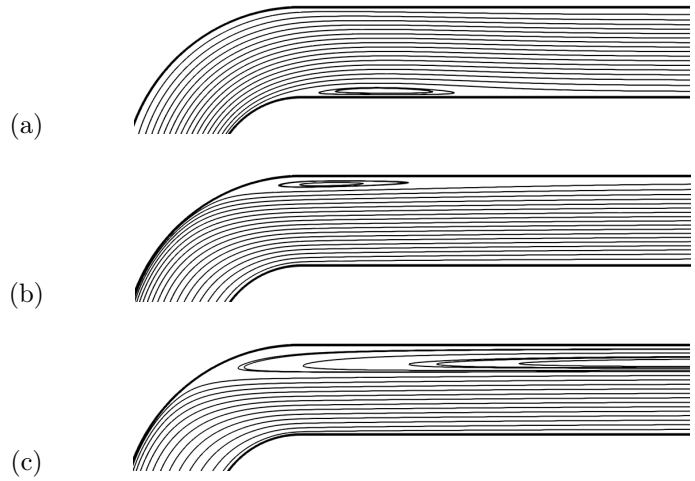


Fig. 8. Streamlines at $Re = 1000$, $R = 2$: $Ha = 10$ (a), $Ha = 65$ (b), $Ha = 100$ (c)

problems, and the identification of general laws of magnetohydrodynamic flows in bends, would have a great influence on the design of blankets and the interpretation of experimental results.

References

[1] P.A.Davidson, Magnetic damping of jets and vortices, *Journal of Fluid Mechanics*, **299**(1995), 153–186. DOI: 10.1017/S0022112095003466

[2] P.A.Davidson, The role of angular momentum in the magnetic damping of turbulence, *Journal of fluid mechanics*, **336**(1997), 123–150. DOI: 10.1017/S002211209600465X

[3] A.V.Proskurin and Anatoly M Sagalakov, An origin of magnetohydrodynamic reverse flow in 90 degree bends, *Physics of Fluids*, **30**(2018), no. 8, 081701. DOI: 10.1063/1.5046328

[4] D.Lee, H.Choi, Magnetohydrodynamic turbulent flow in a channel at low magnetic Reynolds number, *Journal of Fluid Mechanics*, **439**(2001), 367–394.
DOI: 10.1017/S0022112001004621

[5] A.V.Proskurin, A.M.Sagalakov, Spectral/hp element MHD solver, *Magnetohydrodynamics*, **54**(2018), no. 4, 361–372.

Режимы магнитогидродинамического течения в изогнутом канале

Александр В. Проскурин
Алтайский государственный технический университет
Барнаул, Российская Федерация
Анатолий М. Сагалаков
Алтайский государственный университет
Барнаул, Российская Федерация

Аннотация. В работе рассмотрены режимы течения электропроводящей жидкости в изогнутом на 90 градусов канале. Магнитное поле направлено параллельно выходному патрубку канала. Использовались уравнения магнитной гидродинамики в приближении малых магнитных чисел Рейнольдса и спектрально-элементный метод. Паттерны течения изучены при разных значениях чисел Рейнольдса и Гартмана, разных радиусах изгиба канала. Обнаружено возникновение противотечения в выходной части канала.

Ключевые слова: магнитная гидродинамика, течения в каналах, спектрально-элементный метод

DOI: 10.17516/1997-1397-2020-13-6-781-791
УДК 519.6

Accuracy of Symmetric Multi-Step Methods for the Numerical Modelling of Satellite Motion

Evgenia D. Karepova*
Institute of computational modeling of SB RAS
Krasnoyarsk, Russian Federation
Iliya R. Adaev†
Institute of computational modeling of SB RAS
Krasnoyarsk, Russian Federation
Siberian Federal University
Krasnoyarsk, Russian Federation
Yury V. Shan'ko‡
Institute of computational modeling of SB RAS
Krasnoyarsk, Russian Federation

Received 10.08.2020, received in revised form 08.09.2020, accepted 07.10.2020

Abstract. Stability of high-order linear multistep Störmer-Cowell and symmetric methods are discussed in detail in this paper. Efficient algorithms for obtaining intervals of absolute stability and periodicity are given for these methods. To demonstrate the accuracy of numerical integration of the orbit over an interval about one year two model problems are considered. First problem is the 3D Kepler problem. Second one is a specially designed 3D model problem that has the exact solution and simulates the Earth-Moon-satellite system.

Keywords: linear multistep method, symmetric method, Störmer-Cowell method, PECE scheme, orbit.

Citation: E.D. Karepova, I.R. Adaev, Y.V. Shan'ko, Accuracy of the Symmetric Multi-Step Methods for the Numerical Modelling of Satellite Motion, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 781–791. DOI: 10.17516/1997-1397-2020-13-6-781-791.

Introduction

Accuracy of the numerical integration of a satellite motion still remains one of the top problems associated with Global Navigation Satellite Systems. A review of the approaches used by Analysis centres of International GNSS Service [1] shows that the basic techniques of the numerical integration of a satellite orbit are the Adams-Bashforth/Moulton PECE-algorithms, the nonlinear Everhart's procedure [2] and collocation methods [3, 4]. However, a linear multi-step symmetric methods shows considerable promise [5] for near-circular orbits that are typical for navigation satellites.

The theory of multi-step methods, including the Adams family which are traditional for the numerical integration of the motion of celestial objects, are widely discussed in many textbooks

*e.d.karepova@icm.krasn.ru https://orcid.org/0000-0002-6515-2932
†adaev@icm.krasn.ru https://orcid.org/0000-0002-5670-3747
‡shy70@mail.ru https://orcid.org/0000-0003-2796-4363

on numerical methods [6,7,9–12]. The Störmer-Cowell methods were developed and successfully used since the early 20th century. However, in 2016 an interesting result concerning instability for small step size of some Störmer-Cowell methods was presented by Nørsett and Asheim [13]. The general theory of the symmetric multi-step methods was developed by Lambert and Watson [14]. The symmetric methods of high order were discussed in relation to the numerical integration of planetary orbits over a long period of time.

The orbital motion is described by the system of second order ordinary differential equations (ODE). It is generally agreed that is better to solve numerically the second order ODE rather than equivalent system of two first order equations [6,15]. We also confirm this in our numerical experiments.

In this paper, we discuss the accuracy and stability of high-order explicit symmetric multi-step methods and their advantage over the Störmer-Cowell methods with/without "predict – evaluate – correct evaluate" (PECE) mode. We propose an efficient way to calculate intervals of absolute stability and periodicity for any linear multi-step method.

To study stability and periodicity we used the general-purpose computer algebra system REDUCE over the complex field with an accuracy of 40 significant digits. Numerical algorithms were implemented in C++ using the library `quadmath` for quadruple precision calculations.

1. Linear multistep methods

On the discrete point set $\{t_n : \ t_n = t_0 + nh, \ h > 0, n = 0, 1, \dots\}$, we consider the k-step linear multistep method

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = h^2 \sum_{j=0}^{k} \beta_j f_{n+j}, \quad k \geqslant 2, \tag{1}$$

for the numerical solution of the special second-order initial value problem

$$x'' = f(t, x), \quad x(t_0) = x_0, \quad x'(t_0) = \hat{x}. \tag{2}$$

Here x_n is the approximation of the exact solution $x(t_n) \in \mathbb{R}$ and $f_n = f(t_n, x_n)$. Method (1) is characterized by polynomials $\rho(\xi)$ and $\sigma(\xi)$, where

$$\rho(\xi) = \sum_{j=0}^{k} \alpha_j \xi^j, \qquad \sigma(\xi) = \sum_{j=0}^{k} \beta_j \xi^j, \qquad \xi \in \mathbb{C}.$$

We suppose that ρ and σ have no common factors, $\alpha_k = 1$, $|\alpha_0| + |\beta_0| \neq 0$, and $\sum\limits_{j=0}^{k} |\beta_j| \neq 0$. If $\beta_k = 0$ the method is explicit, otherwise it is implicit. For method (1) to be *consistent*, it is necessary and sufficient that $\rho(1) = \rho'(1) = 0$ and $\rho''(1) = 2\sigma(1)$. Method (1) has the order p if for all sufficiently smooth test functions $z(t)$

$$\sum_{j=0}^{k} \alpha_j z(t + jh) - h^2 \sum_{j=0}^{k} \beta_j z''(t + jh) = C_{p+2} h^{p+2} z^{(p+2)}(t) + \mathcal{O}(h^{p+3}).$$

We assume that if the Cauchy problem (2) is solved with the use of method (1) the accuracy of first starting values x_n, $n = 0, \dots, k-1$ is at least not less than the order of the method.

All Störmer-Cowell methods have $\rho(\xi) = \xi^k - 2\xi^{k-1} + \xi^{k-2}$. Method (1) is *symmetric* if $\alpha_j = \alpha_{k-j}$, $\beta_j = \beta_{k-j}$, $j = 0, \dots, k$. A symmetric method has only even order [7]. We study higher order methods, namely, from 6th to 12th order Störmer-Cowell methods and even order

symmetric methods. Coefficients α_j and β_j for these methods are presented in [13] and [5, 14], respectively.

These methods are consistent and zero-stable. Hence they are convergent [6, 14] and polynomial ρ has the root of multiplicity two at +1. Let us denote the roots of ρ by ξ_s, $s = 1, \ldots, k$, where $\xi_1 = \xi_2 = 1$ are the *principal* roots and the remaining $k-2$ roots are *spurious*. All spurious roots of any Störmer-Cowell method are zero.

We demonstrate main differences between symmetric and Störmer-Cowell methods by the example of the harmonic oscillator equation

$$x'' = -\lambda^2 x, \quad x(t_0) = x_0, \quad x'(t_0) = \hat{x}, \quad \lambda \in \mathbb{R}. \tag{3}$$

That has general solution $x(t) = A\cos\lambda t + B\sin\lambda t$ with period $T = 2\pi/\lambda$.

Using method (1) to solve (3), we obtain the difference equation

$$\sum_{j=0}^{k} \left(\alpha_j + H^2\beta_j\right) x_{n+j} = 0 \tag{4}$$

with general solution

$$x_n = D_1 r_1^n + D_2 r_2^n + \sum_{s=3}^{k} D_s r_s^n. \tag{5}$$

Here $H = \lambda h$, $D_s \in \mathbb{C}$ are constant. Let us assume that all the roots r_s, s=1,. . . ,k of the stability polynomial

$$\pi(r; H^2) = \rho(r) + H^2\sigma^2(r) \tag{6}$$

are distinct. Since the roots of the polynomial are continuous functions of its coefficients, r_s are perturbation of ξ_s when $H^2 > 0$. Thus, the numerical solution of (3) x_n may be represented by the sum of the component $(x_n)_P = D_1 r_1^n + D_2 r_2^n$ associated with the perturbation of the principal roots and $(x_n)_S$ that arises from perturbation of spurious roots.

Absolute stability of the Störmer and Cowell methods. Root-locus curves for some Störmer and Cowell methods are shown in Fig. 1 (a–i). They are constructed by the "boundary locus" method [12] which gives a general shape of the boundary $|r| = |\exp(i\varphi)| = 1$ of the open stability region in the complex plane H^2. The stability region is always at the left of the curve when we move along the curve as φ increases from 0 to 2π. For example, there is no stability region for the 10th order Störmer's method (Fig. 1 c). Moreover, the stability region near the interval of absolute stability is shown in more detail in Figs. 1 (b, f, h) for methods that are used in our numerical experiments. Let us note that in the general case $\lambda \in \mathbb{C}$ the stability region is determined, while for the harmonic oscillator we obtain the stability interval on the real axis.

In order to determine the stability interval more accurately the Routh-Hurwitz criterion [16] can be used. In this case, a transformation of the region $|r| \leqslant 1$ into the region $\mathrm{Re}(z) \leqslant 0$ is required. There are the Schur-Cohn [12] and the Jury [17–19] criteria that test the strong stability of $\pi(r; H^2)$ directly. The Schur-Cohn and the Jury criteria are convenient for program implementation and they are easily tested for a given H^2.

According to the Jury criterion, the problem of determining the set of all values of H^2 that all roots of $\pi(r; H^2)$ are inside of the unit circle, is reduced to solving the system of k inequalities, where k is the degree of $\pi(r; H^2)$. The left-hand side of each inequality is the ratio of polynomials in H^2 and the right-hand side is zero. The polynomial coefficients are obtained from the coefficients of $\pi(r; H^2)$. Even for small k the system of the inequalities can be analysed analytically only in some cases. For high order methods this task becomes computationally

Table 1. Stability intervals of H^2 for Störmer's and Cowell methods and the interval of periodicity of H^2 for symmetric methods

Order	5	6	7	8
Störmer	$\left(\frac{360}{323}; \frac{60}{49}\right)$	unstable	$(0; 0.3820447\dots)$	$\left(0; \frac{27}{128}\right)$
Cowell	$\left(0; \frac{60}{11}\right)$	$\left(0; \frac{60}{13}\right)$	$\left(\frac{87280}{308407}; \frac{189}{52}\right)$	$\left(\frac{4221504}{1824647}; \frac{189}{71}\right)$
Order	9	10	11	12
Störmer	unstable	unstable	$\left(0; \frac{51975}{1686934}\right)$	$\left(0; \frac{9450}{595163}\right)$
Cowell	$(0; 0.3597184\dots)$	$(0; 1.0218233\dots)$	$\left(0.1898631; \frac{20790}{28687}\right)$	unstable
Order	6	8	10	12
Symmetric	$(0; 0.8021734\dots)$	$(0; 0.5157665\dots)$	$(0; 0.1724269\dots)$	$(0; 0.0456343\dots)$

intensive. For example, when the Jury criterion is applied to the 8th order Störmer method the maximum degree of the polynomial equals to 12, and for the 8th order Cowell method it equals to 117!

We propose the following effective method to determine the boundaries of the stability interval of method (1). We have to define all H^2 for which the polynomial $\pi(r; H^2)$ has a root that belongs to the unit circle. Consider the roots $r^* = \exp(i\varphi)$ and $\overline{r^*} = \exp(-i\varphi)$ of (6), $0 < \varphi < \pi$. Let us represent $\pi(r; H^2)$ in the form:

$$\pi(r; H^2) = S(r; H^2)(r^2 - 2r\cos\varphi + 1) + R(r; H^2)$$

where $S(r; H^2)$ is a polynomial of the order $(k-2)$ in r with real coefficients, $R(r; H^2) = = a_0(H^2, \cos\varphi) + a_1(H^2, \cos\varphi)r$, a_0, $a_1 \in \mathbb{R}$. Since r^* and $\overline{r^*}$ are the roots of both polynomials $\pi(r; H^2)$ and $r^2 - 2r\cos\varphi + 1$, $R(r; H^2) = 0$. Therefore $a_0(H^2, \cos\varphi) = 0$ and $a_1(H^2, \cos\varphi) = 0$. Consider solutions (H_*^2, φ_*) of the last two equations, where $-1 < \cos\varphi_* < 1$. In addition, the case $\varphi = 0$ gives $H_*^2 = 0$ and the case $\varphi = \pi$ immediately gives $H_*^2 = -\sigma(-1)/\rho(-1)$. Choose all $H_*^2 \in \mathbb{R}^+$ only, and they divide $\mathbb{R}^+$ into disjoint intervals. We test polynomial (6) using the Jury criterion for strong stability for some value of $\hat{H}^2$ belonging to each interval. The interval in $\mathbb{R}^+$ for which $\pi(r; \hat{H}^2)$ is strongly stable corresponds to the interval of absolute stability of method (1).

Tab. 1 presents the absolute stability intervals for the Störmer and Cowell methods of orders from 5 to 12 . The results show that not all methods are stable at small H^2. For example, the Cowell method of order 8 has a very short stability interval separated from zero. The presented results are the same as those from [13], with the exception of the 7th order Störmer method for which one more root was found. It is close to but it does not agree with that found in [13]. This reduces the stability interval. In addition, rational boundaries of the stability intervals can be found with our approach find if they exist.

Interval of periodicity of symmetric methods. If $\hat{\xi}$ is a root of a symmetric polynomial then $1/\hat{\xi}$ is also its root. Then for symmetric method (1) there is no such H^2 that all roots of the stability polynomial $\pi(r; H^2)$ are in the unit circle. Therefore, any symmetric method is absolutely unstable. On the other hand, symmetric methods can have another useful property, namely, they can have a non-vanishing interval of periodicity [14].

According to [14] method (1) has non-vanishing interval of periodicity $(0; H_0^2)$ if for all $H^2 \in (0; H_0^2)$ the roots r_s of the stability polynomial $\pi(r; H^2)$ satisfy relations

$$r_1 = \exp\left(i\theta(H)\right), \quad r_2 = \exp\left(-i\theta(H)\right), \quad |r_s| = 1, \quad s = 3, \ldots, k, \quad \theta(H) \in \mathbb{R}$$

and if the order of (1) is p then $\theta(H) = H + \mathcal{O}(h^{p+1}) \in \mathbb{R}$.

If method (1) has non-vanishing interval of periodicity then it is symmetric. The opposite is not true, but if polynomial ρ of symmetric method (1) has all roots in the unit circle and there are no other double roots except the principal ones then the method has a non-vanishing interval of periodicity. In this case, since the roots of the polynomial continuously depend on parameter H^2, all roots of $\pi(r; H^2)$ remain in the unit circle when H^2 changes from 0 to some H_0^2. Then the principal component $(x_n)_P$ of the numerical solution is periodic with a period close to $2\pi/\lambda$ (the period of the analytical solution of (3)), and $(x_n)_P$ dominates over $(x_n)_S$ which is also periodic.

The approach to determine the value of H_0 is proposed [14]. Some polynomial is constructed from $\pi(r; H^2)$ by special transformation of variable r [14]. The value of H_0^2 is determined from the condition that all roots of the polynomial are real, distinct and non-negative. This corresponds to the condition that the absolute values of all roots of $\pi(r; H^2)$ are equal to 1 for $H^2 \in (0; H_0^2)$.

We propose an alternative method based on determining of H_0^2 in such a way that multiple root arises for $\pi(r; H_0^2)$. Let symmetric method (1) has a non-vanishing interval of periodicity $(0; H_0^2)>$ Then for $H \in (0; H_0^2)$ all roots of $\pi(r; H^2)$ are distinct and lie on the unit circle. Moreover, each root that does not lie on the real axis has the conjugate root as the root of a polynomial with real coefficients (Fig. 2 a). If $H^2 > H_0^2$ then there exists $\xi^* = r^*(\cos\theta^* + i\sin\theta^*)$ root of $\pi(r; H^2)$, where $r^* > 1$. Therefore $1/\xi^* = (\cos\theta^* - i\sin\theta^*)/r^*$ and its conjugate $\overline{\xi^*} = r^*(\cos\theta^* - i\sin\theta^*)$, $\overline{1/\xi^*} = (\cos\theta^* + i\sin\theta^*)/r^*$ are also the roots of $\pi(r; H^2)$. Because r^* is continuously depends on parameter H there exists $H = H^*$ for which $r^* = 1$, that is, ξ^* is root of multiplicity 2. Therefore, H^* coincides with the right-hand boundary of the interval of periodicity H_0. Thus, H_0 can be found as the minimum positive real root of the discriminant of the stability polynomial of a symmetric method. In Fig. 2, the behaviour of the roots of the stability polynomial for the 8th order symmetric method is shown when H approaches H_0 and when H is greater than H_0. Table 1 shows the interval of periodicity for the symmetric methods considered here.

The Störmer methods have a non-vanishing absolute stability interval but do not have an interval of periodicity. Alternatively, symmetric methods are absolutely unstable but they have a non-vanishing interval of periodicity. These differences are shown in Fig. 3 for the following simple numerical example. Let us consider problem (3) with the initial conditions $x(0) = 1$ and $x'(0) = 0$. Then the exact solution is $x(t) = \cos(\lambda t)$. Equation (3) has two the first integrals

$$E := \lambda^2(x(t))^2 + (x'(t))^2 = \texttt{const}, \quad \theta := \lambda t + \arctan\frac{x'(t)}{\lambda x(t)} = \texttt{const}.$$

Although the velocity $x'(t) = v(t)$ is not directly defined by (3), it can be determined by equation (3) through introduction of unknown function v with the initial data $v(0) = 0$, $v'(0) = x''(0) = -\lambda^2$.

The error of the first integral $\Delta E = E^h - E$ for (3) is shown in Fig. 3 (a) and (b) for $\lambda = 1$. equation (3) is integrated with the 8th order symmetric method and Störmer method, respectively. The error of the first integral $\Delta\theta = \theta^h - \theta$ is demonstrated in Fig. 3 (c) for both methods. Here E, θ are exact values of the first integrals (they equal to 1 and 0, respectively) and E^h, θ^h are numerical values of the first integrals. The step-size $h = \pi/128$ belongs to the

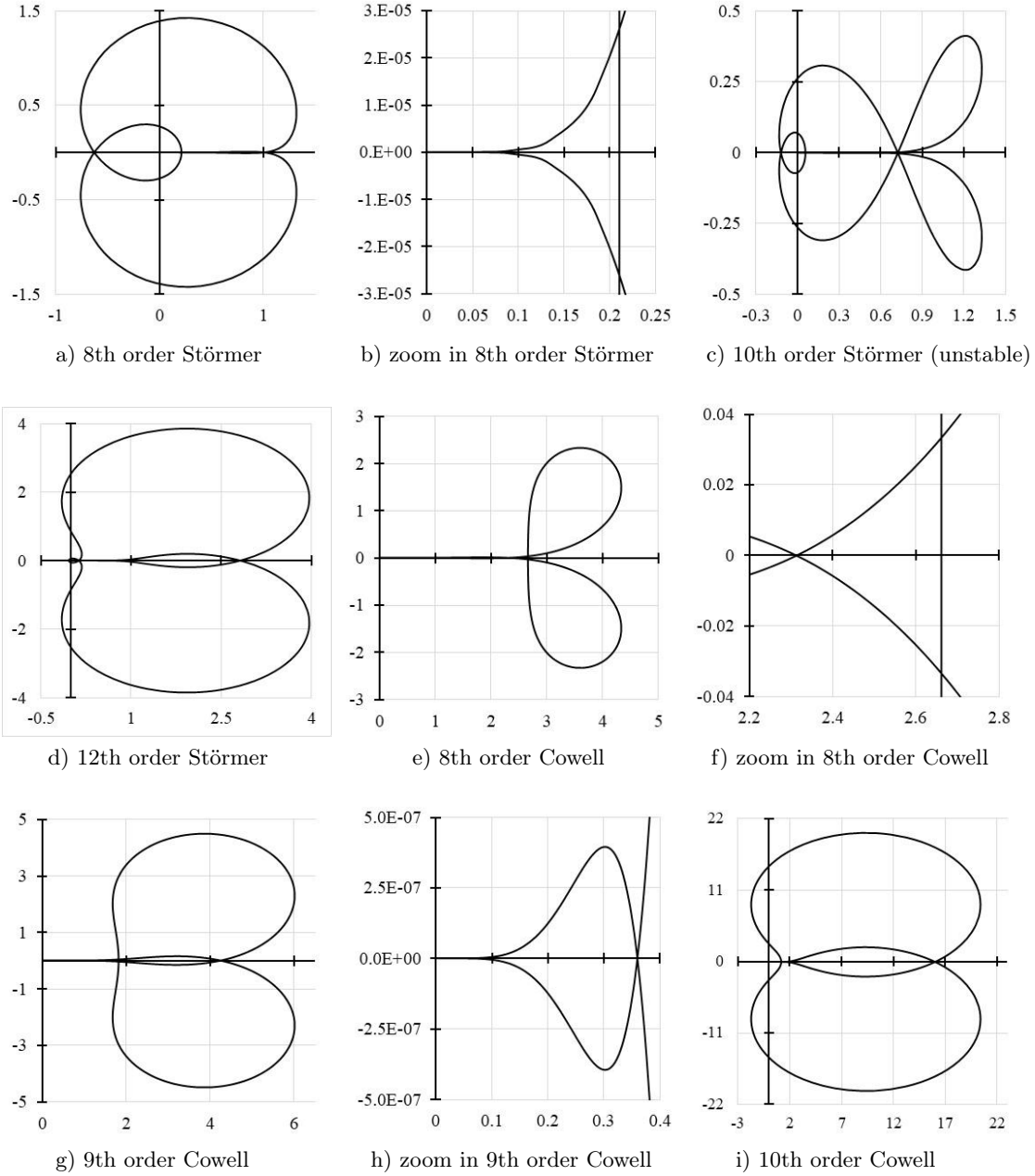
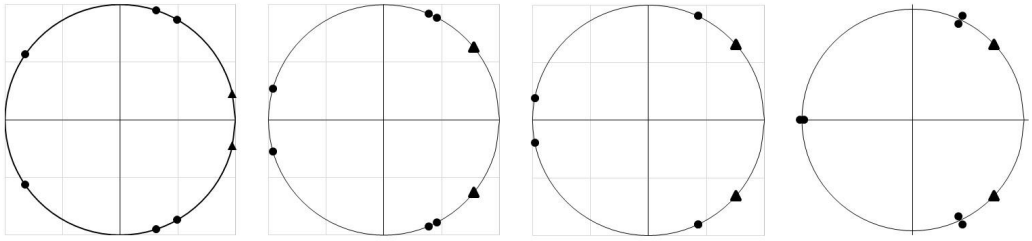


Fig. 1. The root-locus curves and the stability regions for some Störmer and Cowell methods in the complex plane represented by H^2

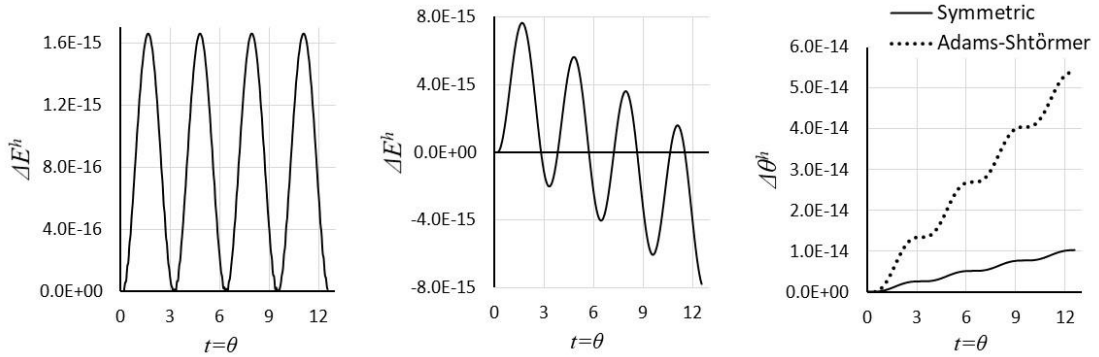
interval of periodicity of symmetric method and to absolute stability interval of the Störmer method.

The symmetric method gives a periodic solution, therefore E^h is a periodic function with constant amplitude. One can see in Fig. 3 (a) that the energy of the system does not increase with time. Since all roots of $\pi(r; H^2)$ for the Störmer method are less than 1, the energy of the numerical solution decreases (Fig. 3 (b)). Since the period of the numerical solution does not



a) b) c) d)

Fig. 2. The roots of stability polynomial $\pi(r; H^2)$ for the 8th order symmetric method (1) in the complex plane; r is shown for $H^2 = H_0^2/10$ (a), $H^2 = 9H_0^2/10$ (b), $H^2 = H_0^2$ (c), $H^2 = 11H_0^2/10$ (d). The roots r_1 and r_2 that correspond to the perturbed principal roots $\xi_1 = \xi_2 = 1$ are marked with black triangle marker



a) b) c)

Fig. 3. The errors of the first integrals for the equation of harmonic oscillator for the symmetric (a,c) and Störmer (b,c) methods with step-size $h/T = 256$, t in radians.

coincide with the theoretical one, the numerical solution is either ahead of or lagging behind the exact solution. For the symmetric method $|\Delta\theta|$ grows slower than for the Störmer method (Fig. 3 (c)).

2. Numerical experiments

Let us consider two three-dimensional model problems that have exact solutions. By "exact solution" is meant a solution that can be obtained by integrating Kepler's equation.

Model problem 1 is the three-dimensional Kepler problem

$$\mathbf{x}''(t) = -\mu \frac{\mathbf{x}}{|\mathbf{x}|^3}, \tag{7}$$

where μ is the standard gravitational parameter, $\mathbf{x} = (x_1, x_2, x_3)$ is the radius-vector of the satellite and $|\mathbf{x}|$ is the Euclidean norm of $\mathbf{x}$.

Model problem 2 is specially constructed from the restricted three-body problem (Earth–Moon–Earth's satellite of negligible mass). In this problem, the force acting from the Moon on the satellite is compensated by an additional force which depends only on time and it is independent of the position of the satellite on the orbit. This force affects the movement of the

satellite in such a way that the exact solution of the problem describes the movement of the satellite around the Earth in the absence of the Moon.

Let us consider the equation of motion for a satellite of negligible mass in an inertial reference frame centred at the Earth-Moon barycentre

$$\mathbf{x}''(t) = -\mu_E \frac{\mathbf{x}-\mathbf{x}_E}{|\mathbf{x}-\mathbf{x}_E|^3} - \mu_M \frac{\mathbf{x}-\mathbf{x}_M}{|\mathbf{x}-\mathbf{x}_M|^3} + \mathbf{f}(t). \tag{8}$$

Here $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{x}_E = ((x_E)_1, (x_E)_2, (x_E)_3)$ and $\mathbf{x}_M = ((x_M)_1, (x_M)_2, (x_M)_3)$ are positions of the satellite, the Earth and the Moon, respectively; μ_E=3.986004419E+14 m^3/s^2 and μ_M=4.9048696E+12 m^3/s^2 are the standard gravitational parameters of the Earth and the Moon; $\mathbf{f}(t) = (f_1(t), f_2(t), f_3(t))$ is an additional force. The coordinates of the Earth and the Moon are determined by two-body problem. Let $\mathbf{x}_{ES}$ be the exact solution of the Kepler problem (7) for the system Earth-satellite. Then $\mathbf{x} = \mathbf{x}_{ES} + \mathbf{x}_E$ and

$$\mathbf{f}(t) = \mu_M \frac{\mathbf{x}_{ES}+\mathbf{x}_E-\mathbf{x}_M}{|\mathbf{x}_{ES}+\mathbf{x}_E-\mathbf{x}_M|^3} - \mu_M \frac{\mathbf{x}_E-\mathbf{x}_M}{|\mathbf{x}_E-\mathbf{x}_M|^3}. \tag{9}$$

Thus, model problem (8)–(9) has the exact solution, and the errors of the numerical solution are calculated directly. Since the Jacobian of the problem coincides with the Jacobian of the restricted three-body problem the stability properties of the numerical methods for these problems coincide.

The following initial orbital parameters are adopted in numerical experiments. For the Moon they are $a_M = 3.94748E + 08$ m, $\varepsilon_M = 0.042200$, $\omega_M = 22°8''$, $\Omega_M = 4°40''$, $i_M = 18°31''$, $(M_0)_M$=340°13″. For satellite they are a_{Sat}= 2.5500000004E+07 m, ε_{Sat}= 0.00068, ω_{Sat}= 135.0000214°, Ω_{Sat}= 120°, i_{Sat}= 64.9°, $(M_0)_{Sat}$= 32.6650111°, T_{Sat}= 11h15′44″.

We compare the accuracy of the orbit integration by the Störmer method and symmetric methods. Additionally, the results for the Bashforth method are shown in the case when problems (7) and (8) are represented in the form of six first-order ODEs. To improve the accuracy of the Adams methods the predictor-corrector scheme is also used in the form $P(EC)^3E$, where the right-hand side (E) and the corrector (C) are evaluated three times at each step. Since the absolute stability intervals of the 8th order Störmer and Cowell methods do not coincide, calculations were carried out for the case when the orders of the predictor and corrector coincide, and they are equal to 8 ($P_8(EC_8)^3E$), and for the case of the 9th order corrector ($P_8(EC_9)^3E$).

For each Model problem, we are interested in the maximum deviation of the calculated satellite position from the exact one after integration for about a year. Let us denote the numerical and exact solutions at the moment t_n by $\mathbf{x}_n^h$ and $\mathbf{x}_{ex}(t_n)$, respectively, $n = k, \ldots, K$, $t_K = 779T_{sat}$, T_{sat} is a period of satellite. The following notations for errors are used

$$\mathbf{\Delta}_n^h = \mathbf{x}_n^h - \mathbf{x}_{ex}(t_n), \quad \Delta_i^h = \max_{n=k,\ldots,K} |\left(\mathbf{\Delta}_n^h\right)_i|, \quad \rho^h = \max_{n=k,\ldots,K} |\mathbf{\Delta}_n^h|.$$

In addition, we consider the decomposition of the error vector $\mathbf{\Delta}_n^h$ in terms of the basis vectors associated with the exact ellipse. They are $\mathbf{r}_0(t_n) = \mathbf{x}_{ex}(t_n)/|\mathbf{x}_{ex}(t_n)|$, $\tau_0(t_n) = = \mathbf{v}_{ex}(t_n)/|\mathbf{v}_{ex}(t_n)|$ and $\mathbf{n_0} = \mathbf{r}_0(t_n) \times \boldsymbol{\tau_0}(t_n)$. Then we have

$$\delta_r^h = \max_{n=k,\ldots,K} |\mathbf{r}_0(t_n) \cdot \mathbf{\Delta}_n^h|, \quad \delta_\tau^h = \max_{n=k,\ldots,K} |\tau_0(t_n) \cdot \mathbf{\Delta}_n^h|, \quad \delta_n^h = \max_{n=k,\ldots,K} |\mathbf{n}_0(t_n) \cdot \mathbf{\Delta}_n^h|.$$

Results of calculations with fixed step-size $h = T_{sat}/512$ for the Model problems 1 and 2 In Tabs are presented in 2, 3, respectively. One can see that the direct solving of the second

order ODE is more efficient. Explicit symmetric methods give more accurate results even in comparison with the explicit–implicit PECE algorithms. In Model problem 1 the Störmer-Cowell PECE algorithm offers slight advantage over other algorithms only in the error along the radius. However, symmetric method offers advantage over other algorithms in calculating positions of the satellite. The symmetric algorithm symmetric method offers advantage over other algorithms in the case of Model problem 2.

Table 2. Accuracy of the orbit integration ($h = T_{sat}/512$). Model problem 1

	Bashforth	Bashforth-Moulton $P_8(EC_8)^3E$	Störmer	Störmer-Cowell $P_8(EC_8)^3E$	Störmer-Cowell $P_8(EC_9)^3E$	symmetric
Δ_1^h, m	7.43E-03	1.77E-04	7.06E-04	2.04E-05	8.45E-06	1.61E-06
Δ_2^h, m	1.07E-02	2.55E-04	1.01E-03	2.93E-05	1.21E-05	2.20E-06
Δ_3^h, m	1.08E-02	2.59E-04	1.03E-03	2.98E-05	1.23E-05	2.27E-06
ρ^h, m	1.20E-02	2.86E-04	1.14E-03	3.29E-05	1.36E-05	2.60E-06
δ_r^h, m	4.98E-06	1.06E-07	3.82E-07	1.73E-08	1.30E-08	1.42E-07
δ_τ^h, m	1.20E-02	2.86E-04	1.14E-03	3.29E-05	1.36E-05	2.60E-06
δ_n^h, m	1.77E-24	1.74E-24	1.40E-22	7.03E-23	1.17E-22	5.07E-23

Table 3. Accuracy of the orbit integration ($h = T_{sat}/512$). Model problem 2

	Bashforth	Bashforth-Moulton $P_8(EC_8)^3E$	Störmer	Störmer-Cowell $P_8(EC_8)^3E$	Störmer-Cowell $P_8(EC_9)^3E$	symmetric
Δ_1^h, m	1.36E-01	3.23E-03	1.27E-02	3.62E-04	1.56E-04	8.89E-05
Δ_2^h, m	1.95E-01	4.64E-03	1.82E-02	5.20E-04	2.23E-04	1.28E-04
Δ_3^h, m	1.97E-01	4.70E-03	1.84E-02	5.27E-04	2.26E-04	1.29E-04
ρ^h, m	2.19E-01	5.22E-03	2.05E-02	5.85E-04	2.51E-04	1.43E-04
δ_r^h, m	3.44E-04	8.14E-06	3.20E-05	9.05E-07	4.02E-07	3.70E-07
δ_τ^h, m	2.19E-01	5.22E-03	2.05E-02	5.85E-04	2.51E-04	1.43E-04
δ_n^h, m	7.55E-04	1.80E-05	9.56E-06	2.73E-07	1.17E-07	6.77E-08

Table 4. Accuracy of the orbit integration ($h = T_{sat}/den$). Model problem 1

	Bashforth	Bashforth-Moulton $P_8(EC_8)^3E$	Störmer	Störmer-Cowell $P_8(EC_8)^3E$	Störmer-Cowell $P_8(EC_9)^3E$	symmetric	symmetric
rhp	486876	1286888	377037	1012680	922316	174497	253176
den	625	413	484	325	296	224	325
h, sec	64.8	98.1	83.7	125	137	181	125
H^2	1.01E-04	2.31E-04	1.69E-04	3.74E-04	4.51E-04	7.87E-04	3.74E-04
ρ^h, m	1.99E-03	1.98E-03	1.89E-03	1.98E-03	1.89E-03	1.91E-03	9.80E-05
δ_r^h, m	6.56E-07	7.51E-07	6.86E-07	7.64E-07	1.81E-06	1.07E-04	5.41E-06
δ_τ^h, m	1.99E-03	1.98E-03	1.89E-03	1.98E-03	1.89E-03	1.91E-03	9.80E-05
δ_n^h, m	1.22E-24	1.12E-24	1.02E-22	4.49E-23	5.65E-23	2.41E-23	5.57E-23

Another series of calculations were carried out to determine the step at which the maximum deviation of the numerical solution from the exact one does not exceed 2 mm for a year. The

Table 5. Accuracy of the orbit integration ($h = T_{sat}/den$). Model problem 2

	Bashforth	Bashforth-Moulton $P_8(EC_8)^3E$	Störmer	Störmer-Cowell $P_8(EC_8)^3E$	Störmer-Cowell $P_8(EC_9)^3E$	symmetric	symmetric
rhp	671499	1779216	519594	1402180	1274424	289789	350551
den	862	571	667	450	409	372	450
h, sec	47.0	71.0	60.8	90.1	99.1	109	90.1
H^2	5.31E-05	1.21E-04	8.87E-05	1.95E-04	2.36E-04	2.85E-04	1.95E-04
ρ^h, m	2.00E-03	1.95E-03	1.87E-03	1.88E-03	1.89E-03	1.84E-03	4.02E-04
δ_r^h, m	3.10E-06	3.03E-06	2.91E-06	2.92E-06	3.03E-06	4.75E-06	1.04E-06
δ_τ^h, m	2.00E-03	1.95E-03	1.87E-03	1.88E-03	1.89E-03	1.84E-03	4.02E-04
δ_n^h, m	6.89E-06	6.72E-06	8.75E-07	8.78E-07	8.86E-07	8.69E-07	1.90E-07

results of calculations are presented in Tabs. 4, 5. The first row marked "rhp" shows the number of evaluations of the right-hand side that were required to achieve the accuracy. In the last column, the results are presented for the symmetric method with the step it takes the Störmer-Cowell PECE algorithm to achieve the specified accuracy. The advantage of the symmetric method is obvious, especially for Model problem 2. In addition, the symmetric methods have the lowest number of right-hand side evaluations in comparison with other methods considered.

This work was supported by the Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation in the framework of the establishment and development of regional Censers for Mathematics Research and Education (Agreement no. 075-02-2020-1631).

References

[1] IGS ftp archives, ftp://ftp.igs.org/pub/center/analysis/. Last accessed 4 Aug 2020

[2] E.Everhart, Implicit Single-Sequence Methods for Integrating Orbits, *Celestial Mechanics*, **10**(1974), 35–55.

[3] G.Beutler, Numerische Integration gewöhnlicher Differentialgleichungssysteme: Prinzipien und Algorithmen. *Mitt. Satell., Beobachtungsstn. Zimmerwald*, **23**(1990).

[4] G.Beutler, Methods of Celestial Mechanics I: Physical, Mathematical, and Numerical Principles, Springer-Verlag, Berlin, 2005.

[5] G.Quinlan, S.Tremaine, Symmetric multistep methods for the numerical integration of planetary orbits, *Astron. J.*, **100**(1990), no. 5, 1694–1700.

[6] P.Henrici, Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons, New York, 1969.

[7] J.D.Lambert, Computational Methods in Ordinary Differential Equations, John Wiley and Sons, New York, 1973.

[8] T.Bordovitcina, The modern numerical methods in problems of celestial mechanics, Nauka, Moscow, 1984 (in Russian).

[9] E.Yairer, S.Norsett, G.Wanner, Solving Ordinary Differential Equations, Springer-Verlag, Berlin, 1987.

[10] E.Vergbitckii, Basis of Numerical Methods, Vysshaya shkola, Moscow, 2004 (in Russian).

[11] V.Avdushev, Numerical modeling of orbits, Izdat. NTI, Tomsk, 2010 (in Russian).

[12] J.C.Butcher, Numerical methods for ordinary differential equations, John Wiley and Sons, New York, 2016.

[13] S.Nørsett, A.Asheim Regarding the absolute stability of Störmer-Cowell methods, *Discrete and Continuous Dynamical Systems*, **34**(2014), no. 3, 1131–1146.
DOI: 10.3934/dcds.2014.34.1131

[14] J.D.Lambert, Symmetric Multistep Methods for Periodic Initial Value Problems, *J. Inst. Maths Applics*, **18**(1976), 189–202.

[15] P.Chakravarti, P.Worland, A class of self-starting methods for the numerical solution of $y'' = f(x, y)$, *BIT Numerical Mathematics*, **11**(1971), no 4, 368–383.

[16] A.Hurwitz, On the conditions under which an equation has only roots with negative real parts (English translation by H. G. Bergmann), in Selected Papers on Mathematical Trends in Control Theory, R. Bellman and R. Kalaba Eds., Dover, New York, 1964, 70–82.

[17] E.Jury, J.Blanchard, A stability test for linear discrete systems in table form, *I.R.E. Proc.*, **49**(1961), 1947–1948.

[18] E.Jury A modified stability table for linear discrete systems, *Proc. IEEE*, **53**(1965), 184–185.

[19] E.Jury, Inners and the Stability of Linear Systems, John Wiley and Sons, New York, 1982.

Точность симметричных многошаговых методов численного моделирования движения спутника

Евгения Д. Карепова
Институт вычислительного моделирования СО РАН
Красноярск, Российская Федерация
Илья Р. Адаев
Сибирский федеральный университет
Красноярск, Российская Федерация
Институт вычислительного моделирования СО РАН
Красноярск, Российская Федерация
Юрий В. Шанько
Институт вычислительного моделирования СО РАН
Красноярск, Российская Федерация

Аннотация. В статье мы подробно обсуждаем устойчивость линейных многошаговых симметричных методов высокого порядка в задаче гармонического осциллятора. Приведены эффективные алгоритмы вычисления интервалов абсолютной устойчивости и периодичности. Численные эксперименты демонстрируют точность вычисления орбиты на интервале около одного года для трехмерной задачи Кеплера и для специально разработанной трехмерной тестовой задачи, которая моделирует систему Земля-Луна-спутник и имеет точное решение.

Ключевые слова: линейные многошаговые методы, симметричный метод, методы Адамса-Штермера-Коуэлла, PECE-схема, орбита.

DOI: 10.17516/1997-1397-2020-13-6-792-796
УДК 539.374

New Classes of Solutions of Dynamical Problems of Plasticity

Sergei I. Senashov[*]
Olga V. Gomonova[†]
Irina L. Savostyanova[‡]
Department of Economic Information Systems,
Reshetnev Siberian State University of Science and Technology,
31 Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russia
Olga N. Cherepanova[§]
Department of Mathematical Analysis and Differential Equations,
Siberian Federal University,
Svobodny 79, Krasnoyarsk, 660041, Russia

Received 10.05.2020, received in revised form 10.06.2020, accepted 20.10.2020

Abstract. *Dynamical problems of the theory of plasticity have not been adequately studied. Dynamical problems arise in various fields of science and engineering but the complexity of original differential equations does not allow one to construct new exact solutions and to solve boundary value problems correctly. One-dimensional dynamical problems are studied rather well but two-dimensional problems cause major difficulties associated with nonlinearity of the main equations. Application of symmetries to the equations of plasticity allow one to construct some exact solutions. The best known exact solution is the solution obtained by B.D. Annin. It describes non-steady compression of a plastic layer by two rigid plates. This solution is a linear one in spatial variables but includes various functions of time. Symmetries are also considered in this paper. These symmetries allow transforming exact solutions of steady equations into solutions of non-steady equations. The obtained solution contains 5 arbitrary functions.*

Keywords: differential equation, plasticity, dynamical problem, exact solution, symmetries.
Citation: S.I. Senashov, O.V. Gomonova I.L. Savostyanova, O.N. Cherepanova, New Classes of Solutions of Dynamical Problems of Plasticity, J. Sib. Fed. Univ. Math. Phys., 2020, 13(6), 792–796.
DOI: 10.17516/1997-1397-2020-13-6-792-796.

Introduction

There is an extensive literature on the theory of plasticity. The reason is that problems considered in this theory are very important for various practical applications. These problems arise in the design of machines and technological processes where plastic deformations are present, in various applications to armaments industry (for example, projectile penetration theory, etc.). Contemporary and classical studies deal mainly with static problems. This is not because dynamical problems are not important but because of lack of progress in developing appropriate

[*]sen@sibsau.ru
[†]gomonova@sibsau.ru
[‡]ruppa@inbox.ru
[§]cheronik@mail.ru

methods to solve these problems. The spatial solution of dynamical equations was first obtained by B. D. Annin in 1978 [2]. This solution is linear in spatial variables and contains several arbitrary functions that depend on time. The solution was constructed with the use of group of point symmetries admitted by the system of equations of dynamical theory of plasticity. Later, new exact solutions of some plane dynamical problems were constructed,. They are based on group properties of the equations. New solutions of the dynamical equations are given in [8]. They are based on transformation of steady-state solutions into non-steady solutions.

New classes of exact solutions of dynamical problems of the theory of plasticity are proposed in the paper. They contain 5 arbitrary functions.

1. Problem definition

Let $x = x_1, y = x_2, z = x_3$ is Cartesian coordinate system, $u = v_1, v = v_2, w = v_3$ are components of strain rate vector, e_{ij} are components of strain velocity tensor, σ_{ij} are components of stress tensor. The components of strain velocity tensor and stress tensor satisfy the equations of motion

$$\frac{dv_i}{dt} = \partial_i \sigma_{ij}, \quad i, j = 1, 2, 3. \tag{1}$$

Here $\dfrac{dv_i}{dt} = \partial_t v_i + v_j \partial_j v_i$ is a full or substantial derivative. Einstein summation convention is applied here. Components of the stress deviator tensor and the strain velocity tensor are coaxial

$$\sigma_{ij} - \delta_{ij} p = \lambda e_{ij} = \lambda(\partial_j v_i + \partial_i v_j)/2, \tag{2}$$

where, δ_{ij} is the Kronecker symbol, λ is a non-negative function, $3p = \sigma_{ii}$.

It is assumed that medium is incompressible. Then we have incompressibility equation

$$\partial_i v_i = 0. \tag{3}$$

In addition to system of equations (1)–(3), von Mises yield criterion is used

$$(\sigma_{11} - p)^2 + (\sigma_{22} - p)^2 + (\sigma_{33} - p)^2 + 2(\sigma_{12}^2 + \sigma_{13}^2 + \sigma_{23}^2) = 2k_s^2, \tag{4}$$

where k_s is the shear yield stress.

2. Group properties of the equations of dynamical theory of plasticity

Lie group of point symmetries admitted by equations (1)–(4) is described in [3]. It is generated by the following operators

$$\begin{gathered} X_0 = \partial_i, \quad M = t\partial_t + x_i \partial_{x_i}, \quad S = \varphi(t)\partial_p, \quad T_i = f_i(t)\partial_{x_i} + f_i'(t)\partial_{v_i} - x_i f_i''(t)\partial_p, \\ Z_1 = x_2\partial_{x_3} - x_3\partial_{x_2} + v_2\partial_{v_3} - v_3\partial_{v_2}. \end{gathered} \tag{5}$$

There is no Einstein summation convention in (5). Two more operators Z_2, Z_3 can be obtained from Z_1 by circular permutation of indices. Functions $\varphi(t)$, $f_i(t)$ are arbitrary functions from the class C^∞. Therefore, operators (5) generate an infinite Lie algebra. Derivatives with respect to variable t is designated by primes.

Group properties of differential equations can be used for various purposes. They are most often used to construct invariant solutions – the solutions which do not change with continuous transformations that correspond to the operators of algebra (5). The invariant solutions of the plasticity equations and methods of their construction are described more fully in [2] and in the literature therein. The procedure of deformation of the exact solutions using point symmetries and the reduction of an exact solution into another one in the case of plane steady equations of ideal plasticity were shown [7]. We use the group of point symmetries for transformation of new stationary solutions into new non-stationary ones for the case of three-dimensional plasticity equations. This approach was firstly applied for construction of new solutions in [8].

3. New stationary solution of system (1)–(4)

As system (1)–(4) admits the operator $X_0 = \partial_t$, one can find the invariant solutions of this system that do not depend on the variable t. These solutions can be determined from the system

$$\begin{gathered} v_j\partial_j v_i = \partial_i\sigma_{ij}, \quad \sigma_{ij} - \delta_{ij}p = \lambda e_{ij} = \lambda(\partial_j v_i + \partial_i v_j)/2, \\ \partial_i v_i = 0, \quad (\sigma_{11}-p)^2 + (\sigma_{22}-p)^2 + (\sigma_{33}-p)^2 + 2(\sigma_{12}^2+\sigma_{13}^2+\sigma_{23}^2) = 2k_s^2. \end{gathered} \tag{6}$$

System (6) is simpler than the initial one because it has fewer independent variables. Some of solutions of the system are given in [8]. As far as we know, there are no other solutions of the considered system [1–3]. Let us find an invariant solution of system (6) regarding the one-dimensional subalgebra that admits the operator $\frac{1}{\alpha}\partial_x + \frac{1}{\beta}\partial_y - \frac{2}{\gamma}\partial_z$. This solution has the following form

$$u = Ag(\alpha x+\beta y+\gamma z),\ v = Bg(\alpha x+\beta y+\gamma z),\ w = Cg(\alpha x+\beta y+\gamma z),\ p = F(\alpha x+\beta y+\gamma z). \tag{7}$$

Here $A,\ B,\ C,\ \alpha,\ \beta,\ \gamma$ are arbitrary constants, and functions $g,\ F$ are determined from system (6). One can obtain the following relations between the functions and the constants

$$\begin{gathered} \alpha A + \beta B + \gamma C = 0, \quad F = \frac{1}{2}g^2 + \delta, \\ \alpha A^2 + \beta AB + \gamma AC = \alpha, \quad \alpha AB + \beta B^2 + \gamma BC = \beta, \quad \alpha AB + \beta BC + \gamma C^2 = \gamma, \end{gathered} \tag{8}$$

here δ is an arbitrary constant. Equalities (7) and (8) imply that all components of the stress tensor are constant and have the form

$$\begin{gathered} \sigma_{11} = p + \frac{\alpha A}{D}, \quad \sigma_{22} = p + \frac{\beta B}{D}, \quad \sigma_{33} = p + \frac{\gamma C}{D}, \\ \sigma_{12} = \frac{\beta A + \alpha B}{2D}, \quad \sigma_{13} = \frac{\gamma A + \alpha C}{2D}, \quad \sigma_{23} = \frac{\gamma B + \beta C}{2D}, \\ D^2 = 2k_s^2\left((\alpha A)^2 + (\beta B)^2 + (\gamma C)^2 + \frac{1}{2}(\beta A + \alpha B)^2 + \frac{1}{2}(\gamma A + \alpha C)^2 + \frac{1}{2}(\gamma B + \beta C)^2\right). \end{gathered} \tag{9}$$

The similar solution with the absence of convective terms was constructed in [8].

4. Deformation of stationary solution of system (1)–(4)

Here, the stationary solution obtained above with the use of transformations (5) is deformed into non-stationary solution of initial system (1)–(4). For this purpose, a notable property of

the point symmetries is used, namely, the symmetries transform any exact solution of system (1)–(4) into a new exact solution of this system.

System (1)–(4) admits operators $S = \varphi(t)\partial_p$, $T_i = f_i(t)\partial_{x_i} + f_i'(t)\partial_{v_i} - x_i f_i''(t)\partial_p$, $(i = 1,\ 2,\ 3)$. It means that the system is not changed under the following transformations

$$x_i' = x_i + a_i f_i(t), \quad v_i' = v_i + a_i f_i'(t), \quad p_i' = p - \sum_{i=1}^{3} a_i x_i f_i''(t) + a_4\varphi(t). \tag{10}$$

Here variables without primes are initial ones and variables with primes are obtained as a result of point symmetries that correspond to subalgebra generated by the operators S, T_i. Parameters a_i are group parameters which change continuously in neighbourhood of zero $x_1 = x,\ x_2 = y,\ x_3 = z$.

Let us assume that v_i^1, p^1 is a solution of system (1)–(4). Then, in accordance with (9), v_i^2, p^2 of the form

$$\begin{aligned}
v_1^2 &= v_1^1\Big(t,\ x_1 + a_1 f_1(t),\ x_2 + a_2 f_2(t),\ x_3 + a_3 f_3(t)\Big) + a_1 f_1'(t),\\
v_2^2 &= v_2^1\Big(t,\ x_1 + a_1 f_1(t),\ x_2 + a_2 f_2(t),\ x_3 + a_3 f_3(t)\Big) + a_2 f_2'(t),\\
v_3^2 &= v_3^1\Big(t,\ x_1 + a_1 f_1(t),\ x_2 + a_2 f_2(t),\ x_3 + a_3 f_3(t)\Big) + a_3 f_3'(t),\\
p^2 &= p^1\left(t,\ x_1 + a_1 f_1(t),\ x_2 + a_2 f_2(t),\ x_3 + a_3 f_3(t)\Big) - \sum_{i=1}^{3} x_i f_i''(t)\right)
\end{aligned} \tag{11}$$

are also an exact solution of the same system. This property is used to construct new solutions of system (1)–(4). Let us apply formulae (11) to the solution constructed above. Then we obtain

$$\begin{aligned}
u &= Ag\Big(\alpha(x + a_1 f_1(t)) + \beta(y + a_2 f_2(t)) + \gamma(z + a_3 f_3(t)\Big) + a_1 f_1'(t),\\
v &= Bg\Big(\alpha(x + a_1 f_1(t)) + \beta(y + a_2 f_2(t)) + \gamma(z + a_3 f_3(t)\Big) + a_2 f_2'(t),\\
w &= Cg\Big(\alpha(x + a_1 f_1(t)) + \beta(y + a_2 f_2(t)) + \gamma(z + a_3 f_3(t)\Big) + a_3 f_3'(t),\\
p &= \frac{1}{2}g^2\Big(\alpha(x + a_1 f_1(t)) + \beta(y + a_2 f_2(t)) + \gamma(z + a_3 f_3(t)\Big) -\\
&\quad - x a_1 f_1''(t) - y a_2 f_2''(t) - z a_3 f_3''(t) + \varphi(t).
\end{aligned} \tag{12}$$

The components of the stress tensor corresponded to the velocity field (12) coincide with (9).

Conclusion

A non-steady solution containing 5 variable functions was constructed from a stationary solution. The method of construction of non-stationary solutions of dynamical equations of plasticity from a stationary solution was shown in the paper. These solutions can be used for the analysis of technological processes when the stress state is stationary but the process is dynamical.

This work was supported by the Krasnoyarsk Mathematical Center and financed by the Ministry of Science and Higher Education of the Russian Federation in the framework of the

establishment and development of regional Centers for Mathematics Research and Education (Agreement No. 075-02-2020-1631).

References

[1] Ivlev D. D. et al. The limiting state of deformable bodies and rocks, Moscow, Physmathlit, 1964 (in Russian).

[2] Annin B. D., Bytev V. O., Senashov S. I. Group properties of equations of elasticity and plasticity, Novosibirsk, Nauka, 1985 (in Russian).

[3] Polyanin A.D., Zaittsev V.F. Handbook of nonlinear partial differential equations, CRC Press, London, New York, Second Edition, 2012.

[4] Novatsky V. K. Wave problems of the theory of plasticity. Moscow, Mir, 1978 (in Russian)

[5] Zadoyan M. A. Space problems of the theory of plasticity. Moscow, Nauka, 1992 (in Russian)

[6] Ishlinky A. Yu., Ivlev D. D. Mathematical theory of plasticity. Moscow, Physmathlit, 2001 (in Russian)

[7] Senashov S.I., Yakhno A.N. Reproduction of solutions for bidimensional ideal plasticity, *Journal of Non -Linear Mechanics*, **42**(2007), 500-503.

[8] Senashov S. I., Savostyanova I. L. New solutions of dynamical equations of plasticity, *Journal of Applied and Industrial Mathematics*, XXII, 4(80), 2019, 89-94.

Новые классы решений динамических задач пластичности

Сергей И. Сенашов
Ольга В. Гомонова
Ирина Л. Савостьянова
Сибирский государственный университет науки и технологий им. Решетнева,
Красноярский рабочий 31, Красноярск, 660037, Россия
Ольга Н.Черепанова
Сибирский федеральный университет,
Свободный 79, Красноярск, 660041, Россия

Аннотация. *Динамические задачи – это наименее изученная область теории пластичности. Динамические задачи возникают в самых разных областях техники и науки, но сложность исходных дифференциальных уравнений не позволяет строить точные решения и корректно численно решать краевые задачи. Неплохо исследованы одномерные динамические задачи пластичности, но уже двумерные вызывают непреодолимые математические сложности, вызванные нелинейностью основных уравнений. Изучение симметрий уравнений пластичности позволило построить некоторые точные решения. Наиболее известное из них это решение Б.Д.Аннина, описывающее нестационарное сжатие пластического слоя жесткими плитами. Это решение линейно по пространственным переменным, но в него входят произвольные функции времени. В предлагаемой работе также используются симметрии.*

Ключевые слова: *дифференциальные уравнения, пластичность, динамические задачи, точные решения, симметрии.*