Ammosov North-Eastern Federal University

# YAKUTIAN MATHEMATICAL JOURNAL

PUBLISHED SINCE 2014 SCIENTIFIC JOURNAL 4 ISSUES PER YEAR

**Vol. 22, No. 4 (88)** **October—December, 2015**

## *CONTENTS*

### Mathematics



### Mathematical Modeling



Address for correspondence:

Ammosov North-Eastern Federal University, Belinskii St., 58, Yakutsk, 677000

Phone: 8(4112)32-14-99, Fax: 8(4112)36-43-47;

http://s-vfu.ru/universitet/rukovodstvo-i-struktura/instituty/niim/mzsvfu/

e-mail: prokopevav85@gmail.com; madu@ysu.ru: ivanegorov51@mail.ru

*Yakutian Mathematical Journal*
*October—December, 2015. Vol. 22, No. 4*

*UDC 517.9*

# ON SOLVABILITY OF DEGENERATE LOADED SYSTEMS OF EQUATIONS

**L. V. Borel**

**Abstract.** We study loaded linear systems of differential equations that are not solved with respect to the time derivative. The systems of partial and ordinary differential equations are examined. Sufficient conditions for the unique solvability of initial boundary value problems (or initial problems) are obtained for these systems of equations.

**Keywords:** loaded equation, degenerate evolution equation, integral operator, initial boundary value problem, algebraic-differential system of equations

## Introduction

Let $\mathfrak{U}$ and $\mathfrak{V}$ be Banach spaces. Given linear operators $L : \mathfrak{U} \to \mathfrak{V}$ and $M : D_M \to \mathfrak{V}$, assume that $L$ is continuous (for brevity we write $L \in \mathscr{L}(\mathfrak{U};\mathfrak{V})$) and $M$ is closed and densely defined in $\mathfrak{U}$; i.e., $M \in \mathscr{C}l(\mathfrak{U};\mathfrak{V})$. We consider the Cauchy problem

$$u(0) = u_0, \tag{1}$$

for the integro-differential equation

$$L\dot{u}(t) = Mu(t) + \int\limits_0^T \mathscr{K}(t,s)u(s)\,d\mu(s), \quad t \in [0,T], \tag{2}$$

that is not solved with respect to the time derivative, since we suppose that $\ker L \neq \{0\}$. Here $T > 0$, $\mathscr{K} : [0,T] \times [0,T] \to \mathscr{L}(\mathfrak{U};\mathfrak{V})$, and $\mu : [0,T] \to \mathbb{R}$ is a function of bounded variation. The equations containing a functional of an unknown solution (for example a Stieltjes integral in (2)) together with a differential part are often called *loaded* [1–4] and they are met when we look for an approximate solution to a differential equation or study inverse problems for them and in the mathematical modeling of nonlocal processes.

A function $u \in C^1([0,T];\mathfrak{U}) \cap C([0,T];D_M)$ satisfying (2) on $[0,T]$ as well as (1) is called a *solution to problem* (1), (2). Sufficient conditions for the unique solvability of the Cauchy problem and the generalized Showalter–Sidorov problem for a loaded equation (2) under the condition of the strong $(L,p)$-radiality of the operator $M$ [6–8] are given in [5]. The effectiveness of the results obtained is demonstrated by the examples of boundary value problems for loaded pseudoparabolic equations. We can refer not only to articles devoted to degenerate evolution equations with a Fredholm integral operator as in (2) but also to those with a Volterra integral operator (the equations with memory) [9–11] or an integral delay operator [12, 13].

The author was supported by the Laboratory of Quantum Topology of Chelyabinsk State University (Grant 14.Z50.31.0020).

The aim of this article is to study some loaded system of equations that are not solved with respect to the time derivative which arises in applications with the use of the results of [5].

A brief exposition of the definitions and theorems from [5, 7] is given in Section 1. Sufficient conditions for unique solvability of boundary value problems for a degenerate loaded system of partial differential equations are obtained in Section 2. Similar results are justified in Section 3 for a degenerate loaded system of equations for a function of one variable, i.e. for algebraic-differential systems (see [14]).

## 1. Degenerate Loaded Equations in Banach Spaces

We recall some definitions and theorems employed in the main part of the article from [5, 7] (see also [6, 8]).

Put

$$\mathbb{N}_0 = \{0\} \cup \mathbb{N}, \quad \overline{\mathbb{R}}_+ = \{0\} \cup \mathbb{R}_+,$$
$$\rho^L(M) = \{\mu \in \mathbb{C} : (\mu L - M)^{-1} \in \mathscr{L}(\mathfrak{V}, \mathfrak{U})\},$$
$$R^L_\mu(M) = (\mu L - M)^{-1} L, \quad L^L_\mu = L(\mu L - M)^{-1}.$$

Let $p \in \mathbb{N}_0$. An operator $M$ is *strongly* $(L,p)$*-radial* whenever

(i) $\exists a \in \mathbb{R}$ $(a, +\infty) \subset \rho^L(M)$;

(ii) $\exists K \in \mathbb{R}_+$ $\forall \mu \in (a, +\infty)$ $\forall n \in \mathbb{N}$

$$\max\left\{\left\|\left(R^L_\mu(M)\right)^{n(p+1)}\right\|_{\mathscr{L}(\mathfrak{U})}, \left\|\left(L^L_\mu(M)\right)^{n(p+1)}\right\|_{\mathscr{L}(\mathfrak{V})}\right\} \le \frac{K}{(\mu - a)^{n(p+1)}};$$

(iii) there exists a dense subspace $\overset{\circ}{\mathfrak{V}}$ in $\mathfrak{V}$ such that

$$\left\|M(\mu L - M)^{-1}\left(L^L_\mu(M)\right)^{p+1} f\right\|_{\mathfrak{V}} \le \frac{\mathrm{const}(f)}{(\mu - a)^{p+2}} \quad \forall f \in \overset{\circ}{\mathfrak{V}};$$

$$\left\|\left(R^L_\mu(M)\right)^{p+1}(\mu L - M)^{-1}\right\|_{\mathscr{L}(\mathfrak{V};\mathfrak{U})} \le \frac{K}{(\mu - a)^{p+2}}$$

for every $\mu \in (a, +\infty)$.

Remark 1. The equivalence of (ii) and (iii) to the slightly more complicated conditions from [6–8] is demonstrated in [15].

Assume that $\mathfrak{U}^0 = \ker\left(R^L_\mu(M)\right)^{p+1}$ and $\mathfrak{V}^0 = \ker\left(L^L_\mu(M)\right)^{p+1}$, while $\mathfrak{U}^1$ is the closure of the range $\mathrm{im}\left(R^L_\mu(M)\right)^{p+1}$ in $\mathfrak{U}$, and $\mathfrak{V}^1$ is the closure of the range $\mathrm{im}\left(L^L_\mu(M)\right)^{p+1}$ in $\mathfrak{V}$. Denote by $L_k$ $(M_k)$ the restriction of $L$ $(M)$ to $\mathfrak{U}^k$ $(D_{M_k} = D_M \cap \mathfrak{U}^k)$, $k = 0, 1$.

**Theorem 1** [7]. *Let $M$ be a strongly $(L,p)$-radial operator. Then*

(i) $\mathfrak{U} = \mathfrak{U}^0 \oplus \mathfrak{U}^1$, $\mathfrak{V} = \mathfrak{V}^0 \oplus \mathfrak{V}^1$;

(ii) $L_k \in \mathscr{L}(\mathfrak{U}^k; \mathfrak{V}^k)$, $M_k \in \mathscr{C}l(\mathfrak{U}^k; \mathfrak{V}^k)$, $k = 0, 1$;

(iii) *there exist* $M_0^{-1} \in \mathscr{L}(\mathfrak{V}^0; \mathfrak{U}^0)$ *and* $L_1^{-1} \in \mathscr{L}(\mathfrak{V}^1; \mathfrak{U}^1)$;

(iv) $H = M_0^{-1} L_0$ *is nilpotent of degree at most* $p$;

(v) *there exists a strongly continuous resolving semigroup* $\{U(t) \in \mathscr{L}(\mathfrak{U}) : t \ge 0\}$ $(\{V(t) \in \mathscr{L}(\mathfrak{V}) : t \ge 0\})$ *of the equation* $L\dot{u}(t) = Mu(t)$. *In this case*

$$\forall t > 0 \quad U(t) = s\text{-}\lim_{k\to\infty}\left(\frac{k(p+1)}{t} R^L_{\frac{k(p+1)}{t}}(M)\right)^{k(p+1)},$$
$$\forall t \ge 0 \quad \|U(t)\|_{\mathscr{L}(\mathfrak{U})} \le K e^{at}.$$

The projection along $\mathfrak{U}^0$ onto $\mathfrak{U}^1$ (along $\mathfrak{V}^0$ onto $\mathfrak{V}^1$) is given by the following formula

$$P = U(0) = s\text{-}\lim_{\mu\to+\infty} \left(\mu R^L_\mu(M)\right)^{p+1} \quad \left(Q = s\text{-}\lim_{\mu\to+\infty} \left(\mu L^L_\mu(M)\right)^{p+1}\right).$$

For a strongly $(L,p)$-radial operator $M$, operator-functions $\mathscr{K}$ of class $C^{p+1,0}([0,T]\times[0,T];\mathscr{L}(\mathfrak{U};\mathfrak{V}))$ (continuous and having continuous partial derivatives with respect to the variables of the first argument up to the order $p+1$) and functions $\mu:[0,T]\to\mathbb{R}$ of bounded variation are denoted by

$$\frac{\partial^n\mathscr{K}}{\partial t^n} \equiv \mathscr{K}^{(n)}_t, \quad V^T_0(\mu) \max_{t,s\in[0,T]} \left\|\mathscr{K}^{(n)}_t(t,s)\right\|_{\mathscr{L}(\mathfrak{U};\mathfrak{V})} \equiv K_n(T),$$

$$V^T_0(\mu) \max_{t,s\in[0,T]} s\left\|\mathscr{K}^{(n)}_t(t,s)\right\|_{\mathscr{L}(\mathfrak{U};\mathfrak{V})} \equiv K_{n,1}(T),$$

$$\left\|L_1^{-1}Q\right\|_{\mathscr{L}(\mathfrak{V};\mathfrak{U})} \equiv C_1, \quad \left\|H^kM_0^{-1}(I-Q)\right\|_{\mathscr{L}(\mathfrak{V};\mathfrak{U})} \equiv h_k, \quad k=0,1,\dots,p,$$

$$F(T) = \max\left\{C_1K(T)\sum_{n=0}^{p+1}K_{n,1}(T) + h_0\sum_{n=0}^{p+1}K_n(T), h_1\sum_{n=0}^{p+1}K_n(T),\dots,h_p\sum_{n=0}^{p+1}K_n(T)\right\}.$$

where $T>0$, $n=0,1,\dots,p+1$. Here $V^T_0(\mu)$ is the variation of $\mu$ on the segment $[0,T]$,

$$K(T) = \max\{K, Ke^{aT}\} = \begin{cases} K, & a\le 0,\\ Ke^{aT}, & a>0,\end{cases}$$

with $K$ and $a$ constants from the definition of strong $(L,p)$-radiality. In view of Theorem 1(v) $\|U(t)\|_{\mathscr{L}(\mathfrak{U})} \le K(T)$ for all $t\in[0,T]$.

**Theorem 2** [5]. *Assume that $M$ is a strongly $(L,p)$-radial operator, $\mu:[0,T]\to\mathbb{R}$ is a function of bounded variation, $\mathscr{K}\in C^{p+1,0}([0,T]\times[0,T];\mathscr{L}(\mathfrak{U};\mathfrak{V}))$, $\mathscr{K}^{(n)}_t(0,s)\equiv 0$, $n=0,1,\dots,p$, $u_0\in D_M\cap\mathfrak{U}^1$, $F(T)<1$. Then there exist a unique solution $u\in C^1([0,T];\mathfrak{U})\cap C([0,T];D_M)$ to* (1), (2).

The problem with the Showalter–Sidorov initial condition

$$Pu(0) = u_0. \tag{3}$$

often met for degenerate evolution equation is addressed for (2) in [5].

**Theorem 3** [5]. *Assume that $M$ is a strongly $(L,p)$-radial operator, $\mu:[0,T]\to\mathbb{R}$ is a function of bounded variation, and $\mathscr{K}\in C^{p+1,0}([0,T]\times[0,T];\mathscr{L}(\mathfrak{U};\mathfrak{V}))$, $u_0\in D_M\cap\mathfrak{U}^1$, $F(T)<1$. Then there exists a unique solution $u\in C^1([0,T);\mathfrak{U})\cap C([0,T];D_M)$ to* (2), (3).

## 2. A Degenerate Loaded System of Partial Differential Equations

Consider the initial-boundary value problem

$$z_1(x,0) = z_{10}(x), \quad x\in\Omega, \tag{4}$$

$$z_i(x,t) = 0, \quad (x,t)\in\partial\Omega\times[0,T], \quad i=1,2,3, \tag{5}$$

for the model integro-differential system of equations

$$
\begin{aligned}
z_{1t}(x,t) &= \Delta z_1(x,t) + \sum_{i=1}^{3} \int_0^T k_{1i}(t,s) z_i(x,s)\, d\mu(s), \quad (x,t) \in \Omega \times [0,T], \\
z_{3t}(x,t) &= \Delta z_2(x,t) + \sum_{i=1}^{3} \int_0^T k_{2i}(t,s) z_i(x,s)\, d\mu(s), \quad (x,t) \in \Omega \times [0,T], \\
0 &= \Delta z_3(x,t) + \sum_{i=1}^{3} \int_0^T k_{3i}(t,s) z_i(x,s)\, d\mu(s), \quad (x,t) \in \Omega \times [0,T].
\end{aligned} \tag{6}
$$

Here $\Omega \subset \mathbb{R}^d$ is a bounded domain with smooth boundary $\partial\Omega$ and the functions $z_{10} : \Omega \to \mathbb{R}$, $k_{ji} : [0,T] \times [0,T] \to \mathbb{R}$, $i,j = 1,2,3$, are given.

Put

$$H_0^2(\Omega) = \{v \in H^2(\Omega) : v(x) = 0,\, x \in \partial\Omega\},$$

$$\mathfrak{U} = \mathfrak{V} = (L_2(\Omega))^3, \quad D_M = \left(H_0^2(\Omega)\right)^3,$$

$$L = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} \Delta & 0 & 0 \\ 0 & \Delta & 0 \\ 0 & 0 & \Delta \end{pmatrix},$$

$$\mathscr{K}(t,s) = \begin{pmatrix} k_{11}(t,s) & k_{12}(t,s) & k_{13}(t,s) \\ k_{21}(t,s) & k_{22}(t,s) & k_{23}(t,s) \\ k_{31}(t,s) & k_{32}(t,s) & k_{33}(t,s) \end{pmatrix}$$

for $t,s \in [0,T]$. In this case $u(t) = \mathrm{col}(z_1(\cdot,t), z_2(\cdot,t), z_3(\cdot,t))$. The strong $(L,1)$-radiality of $M$ is demonstrated and the subspaces

$$\mathfrak{U}^0 = \mathfrak{V}^0 = \{0\} \times L_2(\Omega) \times L_2(\Omega), \quad \mathfrak{U}^1 = \mathfrak{V}^1 = L_2(\Omega) \times \{0\} \times \{0\}$$

and the constants $a = 0$ and $K = 1$ are determined in [16]. Hence, $K(T) \equiv 1$, $L_1^{-1} = I$, $C_1 = 1$, and

$$M_0^{-1} = \begin{pmatrix} \Delta^{-1} & 0 \\ 0 & \Delta^{-1} \end{pmatrix}, \quad H = \begin{pmatrix} 0 & \Delta^{-1} \\ 0 & 0 \end{pmatrix},$$

$$M_0^{-1}(I-Q) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Delta^{-1} & 0 \\ 0 & 0 & \Delta^{-1} \end{pmatrix}, \quad HM_0^{-1}(I-Q) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \Delta^{-2} \\ 0 & 0 & 0 \end{pmatrix}$$

$$h_k = \left\| H^k M_0^{-1}(I-Q) \right\| = \frac{1}{|\lambda_1|^{k+1}}, \quad k = 0,1,$$

where $\lambda_1$ is the first eigenvalue (and thus has the least modulus) of the Laplace operator with the Dirichlet boundary conditions,

$$K_n(T) = V_0^T(\mu) \max_{t,s \in [0,T]} \max_{i,j=1,2,3} \left| \frac{\partial^n k_{ij}}{\partial t^n}(t,s) \right|, \quad n = 0,1,2,$$

$$K_{n,1}(T) = V_0^T(\mu) \max_{t,s \in [0,T]} \left\{ s \max_{i,j=1,2,3} \left| \frac{\partial^n k_{ij}}{\partial t^n}(t,s) \right| \right\}, \quad n = 0,1,2.$$

Hence, Theorem 3 ensures the following claim.

**Theorem 4.** *Assume that $z_{10} \in H_0^2(\Omega)$, $\mu : [0,T] \to \mathbb{R}$ is a function of bounded variation, $k_{ij} \in C^{2,0}([0,T]\times[0,T];\mathbb{R})$, $i,j = 1,2,3$, and*

$$V_0^T(\mu)\max\left\{\sum_{n=0}^{2} K_{n,1}(T) + \frac{1}{|\lambda_1|}\sum_{n=0}^{2} K_n(T), \frac{1}{|\lambda_1|^2}\sum_{n=0}^{2} K_n(T)\right\} < 1.$$

*Then there exists a unique solution $z_1, z_2, z_3 \in C^1([0,T];L_2(\Omega)) \cap C\big([0,T];\, H_0^2(\Omega)\big)$ to problem (4)–(6).*

A particular case of this statement is the following result:

**Corollary 1.** *Assume that $z_{10} \in H_0^2(0,\pi)$, $k_{ij} \in C^{2,0}([0,1]\times[0,1];\mathbb{R})$, $i,j = 1,2,3$,*

$$\sum_{n=0}^{2} \max_{t\in[0,1]} \max_{i,j=1,2,3} \left|\frac{\partial^n l_{ij}}{\partial t^n}(t)\right| < 1.$$

*Then there exists a unique solution $z_1, z_2, z_3 \in C^1([0,1];L_2(0,\pi)) \cap C\big([0,T];\, H_0^2(0,\pi)\big)$ to the problem*

$$z_1(x,0) = z_{10}(x), \quad x \in (0,\pi),$$

$$z_i(0,t) = z_i(\pi,t) = 0, \quad t \in [0,1], \quad i = 1,2,3,$$

$$z_{1t}(x,t) = \Delta z_1(x,t) + \sum_{i=1}^{3} l_{1i}(t) z_i(x,1), \quad (x,t) \in (0,\pi)\times[0,1],$$

$$z_{3t}(x,t) = \Delta z_2(x,t) + \sum_{i=1}^{3} l_{2i}(t) z_i(x,1), \quad (x,t) \in (0,\pi)\times[0,1],$$

$$0 = \Delta z_3(x,t) + \sum_{i=1}^{3} l_{3i}(t) z_i(x,1), \quad (x,t) \in (0,\pi)\times[0,1].$$

Proof. Here $H_0^2(0,\pi) = \{v \in L_2(0,\pi) : v(0) = v(\pi) = 0\}$. Consider the previous theorem with $d = 1$, $\Omega = (0,\pi)$, $T = 1$, $\mu \equiv 0$ on $[0,1)$, $\mu(1) = 1$, $k_{ij}(t,s) = l_i(t)$ for $(t,s) \in [0,1]\times[0,1]$. In this case $V_0^1(\mu) = 1$, $\lambda_1 = -1$, and Theorem 4 validates the claim. □

The solvability of the Cauchy problem (5), (6) with the Cauchy data

$$z_i(x,0) = z_{i0}(x), \quad x \in \Omega, \quad i = 1,2,3, \tag{7}$$

is obtained on using Theorem 2 by analogy.

**Theorem 5.** *Assume that $z_{i0} \in H_0^2(\Omega)$, $i = 1,2,3$, $\mu : [0,T] \to \mathbb{R}$ is a function of bounded variation, $k_{ij} \in C^{2,0}([0,T]\times[0,T];\mathbb{R})$, $i,j = 1,2,3$, $k(0,s) \equiv 0$, $\frac{\partial k}{\partial t}(0,s) \equiv 0$ for $s \in [0,T]$, and*

$$V_0^T(\mu)\max\left\{\sum_{n=0}^{2} K_{n,1}(T) + \frac{1}{|\lambda_1|}\sum_{n=0}^{2} K_n(T), \frac{1}{|\lambda_1|^2}\sum_{n=0}^{2} K_n(T)\right\} < 1.$$

*Then there exists a unique solution $z_1, z_2, z_3 \in C^1([0,T];L_2(\Omega))$ to (5)–(7).*

## 3. A Loaded Algebraic-Differential System of Equations

By analogy, we can establish sufficient conditions for solvability of simpler systems of equations furnished with initial conditions for functions of one variable. The systems are assumed to be algebraic-differential, i.e. they are not solved with respect to the vector of derivatives.

Assume that $B$ and $C$ are square matrices of order $d \in \mathbb{N}$, $\operatorname{rang} B = k$, $k \in \{0, 1, \dots, d-1\}$, $K(t,s)$ is a square matrix of order $d \in \mathbb{N}$, depending on two parameters $t, s \in [0,T]$. Consider the Cauchy problem

$$u(0) = u_0 \tag{8}$$

for the following algebraic-differential system of equations for functions of one variable:

$$B\dot{u}(t) = Cu(t) + \int_0^T K(t,s)u(s)\,d\mu(s), \quad t \in [0,T], \tag{9}$$

where $u(t) = \operatorname{col}(u_1(t), u_2(t), \dots, u_d(t))$, $u_0 = \operatorname{col}(u_{10}, u_{20}, \dots u_{d0})$, and $\mu : [0,T] \to \mathbb{R}$ is a function of bounded variation . The problem (8), (9) agrees with (1), (2) if we put $\mathfrak{U} = \mathfrak{V} = \mathbb{R}^d$ and the action of the operators $L$, $M$, and $\mathscr{K}(t,s)$ is identified with multiplication by the matrices $B$, $C$, and $K(t,s)$, respectively.

**Lemma 1** [17, p. 122]. *Assume that there exists $\alpha \in \mathbb{C}$ such that $\det(\alpha B - C) \neq 0$. Then $M$ is a strongly $(L,p)$-radial operator for some $p \in \{0, 1, \dots, d-1\}$.*

In this case the projection $P$ (see [17, pp. 89–90]) can be calculated by the formula

$$P = \frac{1}{2\pi i}\int_\gamma (\lambda B - C)^{-1} B d\lambda$$

on using residue theory. Under the conditions of Lemma 1 1, Theorem 2 implies that if $u_0 \in \operatorname{im} P$, $\mathscr{K} \in C^{p+1,0}([0,T] \times [0,T]; \mathbb{R}^{d\times d})$, $F(T) < 1$, then there exists a unique solution to (8), (9).

For definiteness, for $d = 3$ we consider the problem

$$u_i(0) = u_i, \quad i = 1, 2, 3, \tag{10}$$

$$\begin{aligned} \dot{u}_1(t) &= u_1(t) + \sum_{i=1}^{3}\int_0^T k_{1i}(t,s)u_i(s)\,d\mu(s), \quad t \in [0,T], \\ \dot{u}_3(t) &= u_2(t) + \sum_{i=1}^{3}\int_0^T k_{2i}(t,s)u_i(s)\,d\mu(s), \quad t \in [0,T], \\ 0 &= u_3(t) + \sum_{i=1}^{3}\int_0^T k_{3i}(t,s)u_i(s)\,d\mu(s), \quad t \in [0,T], \end{aligned} \tag{11}$$

close in the form to that of (5)–(7). Arguing as in Section 2, we infer

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad M = I, \quad (\mu L - M)^{-1} = \begin{pmatrix} \frac{1}{\mu-1} & 0 & 0 \\ 0 & -1 & -\mu \\ 0 & 0 & -1 \end{pmatrix},$$

$$R^L_\mu(M) = \begin{pmatrix} \frac{1}{\mu-1} & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \big(R^L_\mu(M)\big)^2 = \big(L^L_\mu(M)\big)^2 = \begin{pmatrix} \frac{1}{(\mu-1)^2} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\big(R^L_\mu(M)\big)^2(\mu L - M)^{-1} = M(\mu L - M)^{-1}\big(L^L_\mu(M)\big)^2 = \begin{pmatrix} \frac{1}{(\mu-1)^3} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Hence, the operator $I$ is strongly $(L,1)$-radial with the constants $a = 1$, $K = 1$, $K(T) = e^T$ and $\mathfrak{U}^0 = \mathfrak{V}^0 = \{0\} \times \mathbb{R} \times \mathbb{R}$, $\mathfrak{U}^1 = \mathfrak{V}^1 = \mathbb{R} \times \{0\} \times \{0\}$, $L_1^{-1} = I$, $C_1 = h_0 = h_1 = 1$.

**Statement 1.** *Assume that $u_{i0} \in \mathbb{R}$, $i = 1,2,3$, $\mu : [0,T] \to \mathbb{R}$ is a function of bounded variation, $k_{ij} \in C^{2,0}([0,T]\times[0,T];\mathbb{R})$, $i,j = 1,2,3$, $k(0,s) \equiv 0$, $\frac{\partial k}{\partial t}(0,s) \equiv 0$ for $s \in [0,T]$, and*

$$V_0^T(\mu)\sum_{n=0}^{2} \max_{t,s\in[0,T]} \max_{i,j=1,2,3} \left| \frac{\partial^n k_{ij}}{\partial t^n}(t,s) \right| < 1.$$

*Then there exists a unique solution $u_1, u_2, u_3 \in C^1([0,T];\mathbb{R})$ to* (10), (11).

## REFERENCES


1. *Nakhushev A. M.* Loaded equations and their applications // Diff. Equ. 1983. V. 19, N 1. P. 74–81.
2. *Dzhenaliev M. T.* To the Theory of Linear Boundary Value Problems for Loaded Differential Equations [in Russian]. Almaty: Komp′yut. Tsentr ITPM, 1995.
3. *Kozhanov A. I.* Nonlinear loaded equations and inverse problems // Comput. Math. Math. Phys. 2004. V. 44, N 4. P. 657–678.
4. *Kozhanov A. I.* A nonlinear loaded parabolic equation and a related inverse problem // Math. Notes. 2004. V. 76, N 6. P. 784–795.
5. *Fedorov V. E. and Borel L. V.* Solvability of loaded linear evolution equations with a degenerate operator at the derivative // St. Petersburg Math. J. 2015. V. 26, N 3. P. 487–497.
6. *Fedorov V. E.* Linear equations of the Sobolev type with relatively $p$-radial operators // Dokl. Math. 1996. V. 54, N 3. P. 883–885.
7. *Fedorov V. E.* Degenerate strongly continuous semigroups of operators // St. Petersburg Math. J. 2001. V. 12, N 3. P. 471–489.
8. *Fedorov V. E.* A generalization of the Hille–Yosida theorem to the case of degenerate semigroups in locally convex spaces // Sib. Math. J. 2005. V. 46, N 2. P. 333–350.
9. *Falaleev M. V. and Orlov S. S.* Integro-differential equations with degeneration in Banach spaces and it's applications in mathematical theory of elasticity // IIGU Ser. Matematika. 2011. V. 4, N 1. P. 118–134.
10. *Falaleev M. V. and Orlov S. S.* Degenerate integro-differential operators in Banach spaces and their applications // Russian Math. (Iz. VUZ). 2011. V. 55, N 10. P. 59–69.
11. *Fedorov V. E. and Borel L. V.* On solvability of degenerate linear evolution equations with memory effects // IIGU Ser. Matematika. 2014. V. 10. P. 106–124.
12. *Fedorov V. E. and Omel′chenko E. A.* Inhomogeneous degenerate Sobolev type equations with delay // Sib. Math. J. 2012. V. 53, N 2. P. 335–344.
13. *Fedorov V. E. and Omel′chenko E. A.* Linear equations of the Sobolev type with integral delay operator // Izv. Vyssh. Uchebn. Zaved. Mat. 2014. N 1. P. 71–81.
14. *Chistyakov V. V. and Shcheglova A. A.* Selected Chapters of the Theory of Differential Algebraic Equations [in Russian]. Novosibirsk: Nauka, 2003.

**15.** *Fedorov V. E.* The pseudoresolvent property and existence conditions of degenerate semigroups of operators // Vestnik Chelyabinsk. Univ. Mat. Mekh. Inform. 2009. V. 20, N 11. P. 12–19.

**16.** *Ruzakova O. A. and Fedorov V. E.* On $\varepsilon$-controllability of linear equations that are not solved with respect to the derivative in Banach spaces // Vychisl. Tekhnologii. 2005. V. 10, N 5. P. 90–102.

**17.** *Sviridyuk G. A. and Fedorov V. E.* Linear Sobolev Type Equations and Degenerate Semigroups of Operators. Utrecht; Boston: VSP, 2003.

*October 1, 2015*

L. V. Borel
Laboratory of Quantum Topology
Chelyabinsk State University, Chelyabinsk, Russia
lidiya904@mail.ru

*Yakutian Mathematical Journal*
*October—December, 2015. Vol. 22, No. 4*

*UDC 514.132*

# THE PSEUDOVOLUME OF A HYPERBOLIC TETRAHEDRON

**E. S. Kudina and A. D. Mednykh**

**Abstract.** Studying the properties of hyperbolic tetrahedra in the three-dimensional Lobachevsky space, we compare two concepts characterizing them: the volume and the pseudovolume defined as the square root of the absolute value of the Gram determinant formed by the edge lengths. In 1877 d'Ovidio conjectured that these two concepts for hyperbolic tetrahedra coincide up to a natural normalization constant. Later the conjuncture turned out false; nevertheless, the asymptotic equality holds for infinitesimal tetrahedra. The classical theorem of Servois asserts that the volume of every Euclidean tetrahedron equals one sixth the product of lengths of two skew edges by the distance and the sine of the angle between them. We show that this theorem remains valid for the pseudovolumes of hyperbolic tetrahedra, but fails for their non-Euclidean volumes. As a corollary, we find a similar situation for Steiner's theorem on the conservation of the Euclidean volume of a tetrahedron under parallel translations of its edges.

**Keywords:** hyperbolic space, volume, pseudovolume, Servois's theorem, Steinitz's theorem, Gram matrix

## 1. Introduction

To calculate the volume of a polyhedron is the classical problem known since Euclid's time and still remaining worthy of interest. The reason is largely that the volume of a three-dimensional manifold is an important geometric invariant. In this article we consider tetrahedra in hyperbolic space. We compare the volume and the pseudovolume of a hyperbolic tetrahedron in the three-dimensional Lobachevsky space. In 1877 the Italian Enrico d'Ovidio conjectured that these two concepts coincide up to a constant. However, it was shown later that they differ, while the asymptotic equality holds only for infinitesimal tetrahedra.

By the classical theorem of Servois [1, p. 98], the volume of every Euclidean tetrahedron equals one sixth the product of the length of its skew edges by the distance and the sine of the angle between them. An important corollary is Steiner's theorem [1, p. 99] asserting that the volume of every Euclidean tetrahedron is preserved under parallel translations of its edge along the line which includes it. This article presents a proof of Servois's theorem for the pseudovolume of hyperbolic tetrahedra and constructs an example demonstrating that the theorem fails for the volume. As a corollary of Servois's theorem for hyperbolic pseudovolume, we obtain Steiner's theorem for the pseudovolume of a hyperbolic tetrahedron.

The authors weer supported by the Russian Foundation for Basic Research (Grants 16–31–50009 and 15–01–07906).

## 2. Preliminaries

As the model of the hyperbolic geometry $\mathbb{H}^3$, consider the upper half-space $R^3 = \{(x, y, t),\ x, y, t \in \mathbb{R},\ t > 0\}$ endowed with the metric $ds^2 = (dx^2 + dy^2 + dt^2)/t^2$. For all necessary formulas of non-Euclidean geometry, see [2]. We need the following well-known properties.

**Theorem 2.1.** *The geodesics in $\mathbb{H}^3$ are the half-lines and half-circles orthogonal to the $xy$ plane.*

**Theorem 2.2.** *The group of orientation-preserving isometries of the Lobachevsky space $\mathbb{H}^3$ coincides with $PSL(2, \mathbb{C})$ consisting of the Möbius transformations of the form $z \to \frac{az+b}{cz+d}$, where $a, b, c, d \in \mathbb{C}$ with $ad - bc = 1$. The action of these mappings extends to $\mathbb{H}^3$ according to Poincaré's formula*

$$z + tj \to [a(z + tj) + b][c(z + tj) + d]^{-1}, \tag{1}$$

*where $j$ with $j^2 = -1$ is the unit quaternion, $z = x + y\,i$, while $z + t\,j$ are identified with the points $(x, y, t)$ of $\mathbb{H}^3$.*

**Theorem 2.3.** *The distance function $\rho$ on $\mathbb{H}^3$ satisfies*

$$\cosh \rho(z_1 + t_1 j, z_2 + t_2 j) = 1 + \frac{|z_1 - z_2|^2 + |t_1 - t_2|^2}{2t_1 t_2}. \tag{2}$$

**Theorem 2.4.** *The differential of the volume of a tetrahedron in the hyperbolic space $\mathbb{H}^3$ satisfies Schläfli's formula*

$$dV = -\frac{1}{2}\sum_k l_k d\theta_k, \tag{3}$$

*where $l_k$ and $\theta_k$ are the length of the edge $k$ and the dihedral angle along the edge.*

## 3. The Main Results

We start with the following preliminary result.

**Lemma 3.1.** *Given four distinct points $z_1$, $z_2$, $z_3$, and $z_4$ on the extended complex plane, there exists a Möbius transformation carrying them to the points $-R$, $R$, $-\frac{e^{i\varphi}}{R}$, and $\frac{e^{i\varphi}}{R}$, where $R > 0$ and $\varphi$ is some real number.*

PROOF. Consider the mapping

$$T(z) = \frac{z - z_1}{z - z_2} : \frac{z_3 - z_1}{z_3 - z_2}$$

and denote the cross-ratio of $z_1$, $z_2$, $z_3$, and $z_4$ by $Q = T(z_4)$. Then $T(z)$ sends $z_1$, $z_2$, $z_3$, and $z_4$ to 0, $\infty$, 1, and $Q$.

Let us find a Möbius mapping $L$ carrying 0, $\infty$, 1, and $Q$ to $-k$, $\frac{1}{k}$, $k$, and $-\frac{1}{k}$ for some complex number $k$.

The appropriate cross-ratios yield

$$Q = \left(\frac{1}{2}\left(k + \frac{1}{k}\right)\right)^2. \tag{4}$$

Solving the resulting equation for $k$, we find one of its roots. The remaining three roots are $-k$, $-\frac{1}{k}$ and $\frac{1}{k}$. We can verify directly that the Möbius mapping

$$L(z) = \frac{-2kz + k(1 + k^2)}{2k^2 z - (1 + k^2)} \tag{5}$$

has the required properties.

Put $k = Re^{i\psi}$. Then the mapping $z \to e^{-i\psi}z$ carries $-k$, $k$, $-\frac{1}{k}$, and $\frac{1}{k}$ to $-R$, $R$, $-\frac{1}{Re^{2i\psi}}$, and $\frac{1}{Re^{2i\psi}}$. Therefore, the mapping $e^{-i\psi} \circ L \circ T$ carries $z_1$, $z_2$, $z_3$, and $z_4$ to $-R, R, -\frac{1}{Re^{2i\psi}}, \frac{1}{Re^{2i\psi}}$. Putting $\varphi = -2\psi$, we obtain the result. $\square$

## 4. Non-Euclidean Versions of Servois's and Steiner's Theorems

In order to obtain these theorems, we have to express the Gram determinant of a hyperbolic tetrahedron in terms of the lengths of its skew edges, as well as the distance and angle between them. To use the most convenient method, realize a hyperbolic tetrahedron in the upper half-space $\mathbb{H}^3 = \{(x, y, t),\ x, y, t \in \mathbb{R},\ t > 0\}$ so that the common perpendicular to the skew edges coincides with the vertical axis $Ot$.

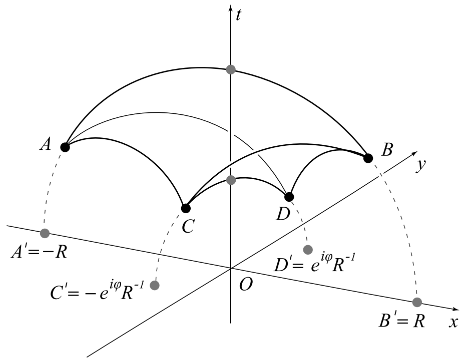



Fig. 1. A hyperbolic tetrahedron $ABCD$

Consider an arbitrary hyperbolic tetrahedron $ABCD$. Extend the side $AB$ from $A$ to a vertex $A'$ lying on the absolute and from $B$ to a vertex $B'$ also lying on the absolute. Similarly extend the side $CD$ from $C$ to $C'$ and from $D$ to $D'$. Applying if need be the Möbius mapping of Lemma 3.1, without loss of generality we may assume that $A' = -R$, $B' = R$, $C' = -\frac{e^{i\varphi}}{R}$, and $D' = \frac{e^{i\varphi}}{R}$.

Denote by $O$ the origin of the Cartesian coordinate system $Oxyt$ and by $\varphi_A$, the angle formed by the rays $OA$ and $OA'$. Similarly, introduce $\varphi_B$, $\varphi_C$, and $\varphi_D$ for the corresponding angles $OBB'$, $OCC'$, and $ODD'$.

Then we can define the vertices of the tetrahedron using Cartesian coordinates:

$$A = (R\cos\varphi_A, 0, R\sin\varphi_A), \quad B = (R\cos\varphi_B, 0, R\sin\varphi_B),$$

$$C = \left(\frac{1}{R}\cos\varphi_C\cos\varphi, \frac{1}{R}\cos\varphi_C\sin\varphi, \frac{1}{R}\sin\varphi_C\right),$$

$$D = \left(\frac{1}{R}\cos\varphi_D\cos\varphi, \frac{1}{R}\cos\varphi_D\sin\varphi, \frac{1}{R}\sin\varphi_D\right).$$

Calculate the hyperbolic distance between the vertices using (2) (see Theorem 2.3):

$$\cosh\rho(A,B) = \frac{1-\cos\varphi_A\cos\varphi_B}{\sin\varphi_A\sin\varphi_B}, \quad \cosh\rho(C,D) = \frac{1-\cos\varphi_C\cos\varphi_D}{\sin\varphi_C\sin\varphi_D},$$

$$\cosh\rho(A,C) = \frac{R^2+R^{-2}-2\cos\varphi\cos\varphi_A\cos\varphi_C}{2\sin\varphi_A\sin\varphi_C},$$

$$\cosh\rho(B,C) = \frac{R^2+R^{-2}-2\cos\varphi\cos\varphi_B\cos\varphi_C}{2\sin\varphi_B\sin\varphi_C},$$

$$\cosh\rho(A,D) = \frac{R^2+R^{-2}-2\cos\varphi\cos\varphi_A\cos\varphi_D}{2\sin\varphi_A\sin\varphi_D},$$

$$\cosh\rho(B,D) = \frac{R^2+R^{-2}-2\cos\varphi\cos\varphi_B\cos\varphi_D}{2\sin\varphi_B\sin\varphi_D}.$$

The Gram matrix of this tetrahedron is

$$G = \begin{pmatrix} -1 & -\cosh\rho(A,B) & -\cosh\rho(A,C) & -\cosh\rho(A,D) \\ -\cosh\rho(A,B) & -1 & -\cosh\rho(B,C) & -\cosh\rho(B,D) \\ -\cosh\rho(A,C) & -\cosh\rho(B,C) & -1 & -\cosh\rho(C,D) \\ -\cosh\rho(A,D) & -\cosh\rho(B,D) & -\cosh\rho(C,D) & -1 \end{pmatrix}.$$

Taking the determinant of $G$, we find

$$\det(G) = -\frac{\sin^2\varphi(R^4-1)^2(\cos\varphi_A-\cos\varphi_B)^2(\cos\varphi_C-\cos\varphi_D)^2}{4R^4\sin^2\varphi_A\sin^2\varphi_B\sin^2\varphi_C\sin^2\varphi_D}. \tag{6}$$

The formula $\sinh^2 z = \cosh^2 z - 1$ yields

$$\sinh\rho(A,B) = \frac{\cos\varphi_A-\cos\varphi_B}{\sin\varphi_A\sin\varphi_B},$$

$$\sinh\rho(C,D) = \frac{\cos\varphi_C-\cos\varphi_D}{\sin\varphi_C\sin\varphi_D}.$$

Thus,

$$\det(G) = -\frac{(R^4-1)^2}{4R^4}\sin^2\varphi\sinh^2\rho(A,B)\sinh^2\rho(C,D). \tag{7}$$

Since $R = e^{\frac{\rho}{2}}$, it follows that $\frac{R^4-1}{2R^2} = \sinh\rho$. Then (7) becomes

$$\det(G) = -\sinh^2\rho\sin^2\varphi\sinh^2\rho(A,B)\sinh^2\rho(C,D). \tag{8}$$

Note that formulas equivalent to (8), although expressed in slightly different geometric terms and proved by completely different technique, were obtained in Fenchel's book [3, p. 169, Eq. (24)] and in the unpublished manuscript of McConnell [4].

Recall that we can calculate the pseudovolume of the hyperbolic tetrahedron in terms of the Gram matrix as $\widetilde{V} = \sqrt{-\det(G)}$. Various authors call this quantity the amplitude, content, polar sine, and so on. Inserting this expression into (8) yields the following analog of Servois's formula for the hyperbolic tetrahedron.

**Theorem 4.1.** *The pseudovolume $\widetilde{V}$ of a hyperbolic tetrahedron, the lengths of its opposite edges $\rho(A,B)$ and $\rho(C,D)$, as well as the angle $\varphi$ and distance $\rho$ between these edges are related as*

$$\widetilde{V} = \sinh\rho\sin\varphi\sinh\rho(A,B)\sinh\rho(C,D). \tag{9}$$

As a corollary to (9), we obtain Steiner's theorem for the hyperbolic tetrahedron.

**Theorem 4.2.** *The pseudovolume of a hyperbolic tetrahedron is preserved when its opposite edges move retaining their lengths along the lines that include them.*

## 5. A Counterexample to Steiner's Theorem

Consider a hyperbolic tetrahedron $OABC$ with three pairwise orthogonal sides meeting at the vertex $O$. Go from $O$ along three pairwise orthogonal geodesics on its edges, which we agree to call the coordinate axes $Ox$, $Oy$, and $Oz$. Assume that the vertices $A$, $B$, and $C$ lie on the corresponding coordinate axes.

Denote by $\cosh x$, $\cosh y$, and $\cosh z$ the hyperbolic cosines of the lengths of the side lying on the corresponding coordinate axes $Ox$, $Oy$, and $Oz$. For convenience, denote the so-constructed tetrahedron (Fig. 2) by $T(x, y, z)$.

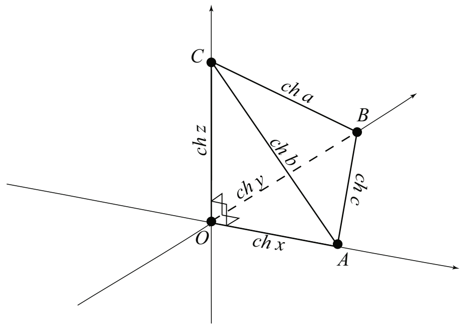



Fig. 2. A hyperbolic tetrahedron $OABC$ with three pairwise orthogonal sides

Pythagoras's theorem for the hyperbolic right triangles $OBC$, $OAC$, and $OAB$ yields

$$\cosh a = \cosh y \cosh z, \quad \cosh b = \cosh x \cosh z, \quad \cosh c = \cosh x \cosh y. \tag{10}$$

The dihedral angles $\alpha$, $\beta$, and $\gamma$ along the edges $BC$, $AC$, and $AB$ of lengths $a$, $b$, and $c$ satisfy [5, p. 130]

$$\begin{aligned}
\cot^2 \alpha &= \frac{(\cosh a \cosh c - \cosh b)(\cosh a \cosh b - \cosh c)}{(\cosh^2 a - 1)(\cosh b \cosh c - \cosh a)}, \\
\cot^2 \beta &= \frac{(\cosh b \cosh c - \cosh a)(\cosh a \cosh b - \cosh c)}{(\cosh^2 b - 1)(\cosh a \cosh c - \cosh b)}, \\
\cot^2 \gamma &= \frac{(\cosh b \cosh c - \cosh a)(\cosh a \cosh c - \cosh b)}{(\cosh^2 c - 1)(\cosh a \cosh b - \cosh c)}.
\end{aligned} \tag{11}$$

Remark. In the notation of [5], we have $\frac{A}{2} = \alpha$, $\frac{B}{2} = \beta$, and $\frac{C}{2} = \gamma$.

Construct the desired counterexample as follows. Put $x = y = z = u$ and consider the tetrahedron $T_1 = T(u, u, u)$ (Fig. 3). On the axis $Oy$ choose two points $D$ and $D'$ symmetric with respect to the plane $OAC$ and lying at the distance $\frac{u}{2}$ from $O$. Denote this tetrahedron $ACDD'$ by $T_2$ (Fig. 4). Then $T_1$ and $T_2$ share the edge $AC$ of length $b$, with the opposite edges $OB$ and $DD'$ of length $u$. Observe that we obtain $DD'$ from the edge $OB$ by a parallel translation along the $Oy$ axis. Therefore, the tetrahedra $T_1$ and $T_2$ satisfy the conditions of Steiner's theorem. Our goal is to show that they have different hyperbolic volumes.

The proof rests on the two lemmas:

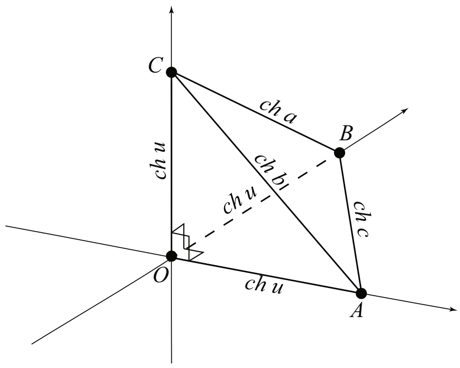



Fig. 3. The tetrahedron $T_1$

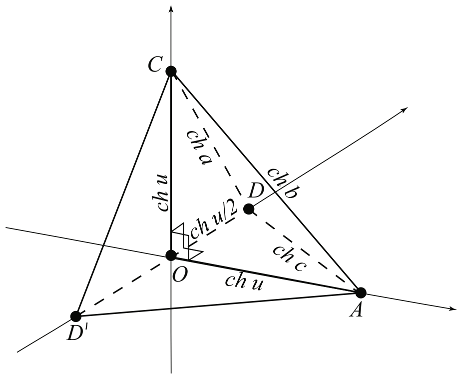



Fig. 4. The tetrahedron $T_2$

**Lemma 5.1.** *The hyperbolic volume* $\mathrm{Vol}(T_1)$ *satisfies*

$$\mathrm{Vol}(T_1) = \int\limits_1^{\cosh u} f(t)\,dt, \tag{12}$$

*where*

$$f(t) = \frac{3\cosh^{-1}(t^2)}{2\sqrt{1+t^2}(1+2t^2)}.$$

**Lemma 5.2.** *The hyperbolic volume* $T_2$ *satisfies*

$$\mathrm{Vol}(T_2) = 2\int\limits_1^{\cosh u} g(t)\,dt, \tag{13}$$

*where*

$$g(t) = \frac{\cosh^{-1}(\frac{t\sqrt{1+t}}{\sqrt{2}})(2+t)}{2(1+t+t^2)\sqrt{2+2t+t^2}} + \frac{\cosh^{-1}(t^2)(1-t)}{4(1+t+t^2)\sqrt{1+t^2}}.$$

Proof of Lemma 5.1. Consider the first tetrahedron $T_1 = T(u,u,u)$ and put $t = \cosh u$. Then $\cosh\frac{u}{2} = \sqrt{\frac{1+t}{2}}$. Find the differentials of the angles $\alpha$, $\beta$, and $\gamma$. By (11),

$$\alpha = \beta = \gamma = \cosh^{-1}\frac{t}{\sqrt{1+t^2}},$$

whence

$$d(\alpha) = d(\beta) = d(\gamma) = -\frac{dt}{(1+2t^2)\sqrt{1+t^2}}. \tag{14}$$

Calculate the lengths of the corresponding sides $l_\alpha = a$, $l_\beta = b$, and $l_\gamma = c$:

$$-\frac{l_\alpha}{2} = -\frac{l_\beta}{2} = -\frac{l_\gamma}{2} = -\frac{1}{2}\cosh^{-1}(t^2). \tag{15}$$

Inserting (14) and (15) into Schläfli's formula (3), we obtain the claim.

Proof of Lemma 5.2. The second tetrahedron $T_2$ consists of two mirror copies of the tetrahedron $T_2' = T\big(u, \frac{u}{2}, u\big)$.

Find the differentials of the angles $\alpha$, $\beta$, and $\gamma$ of $T_2'$. By (11),

$$\alpha = \gamma = \cosh^{-1}\frac{t}{\sqrt{2+2t+t^2}}, \quad \beta = \cosh^{-1}\frac{1+t}{\sqrt{1+t^2}},$$

whence

$$\begin{aligned} d(\alpha) &= -\frac{2+t}{2(1+t+t^2)\sqrt{2+2t+t^2}}dt,\\ d(\beta) &= \frac{t-1}{2(1+t+t^2)\sqrt{1+t^2}}dt, \\ d(\gamma) &= -\frac{2+t}{2(1+t+t^2)\sqrt{2+2t+t^2}}dt. \end{aligned} \tag{16}$$

Calculate the lengths of the corresponding sides:

$$\begin{gathered} -\frac{l_\alpha}{2} = -\frac{1}{2}\cosh^{-1}\left(t\sqrt{\frac{1+t}{2}}\right), \quad -\frac{l_\beta}{2} = -\frac{1}{2}\cosh^{-1}(t^2), \\ -\frac{l_\gamma}{2} = -\frac{1}{2}\cosh^{-1}\left(t\sqrt{\frac{1+t}{2}}\right). \end{gathered} \tag{17}$$

Schläfli's formula yields

$$d\,\mathrm{Vol}(T_2') = -\frac{l_\alpha}{2}d\alpha - \frac{l_\beta}{2}d\beta - \frac{l_\gamma}{2}d\gamma = g(t)\,dt,$$

where $g(t)$ is the same as in the statement of Lemma 5.2. Hence, the obvious equality $\mathrm{Vol}(T_2) = 2\,\mathrm{Vol}(T_2')$ leads to (13).

It is not difficult to see that the functions defined by the integrals (12) and (13) are distinct. Thus, Steiner's theorem fails for the hyperbolic volumes of tetrahedra.

## *REFERENCES*


1. *Ponarin Ya. P.* Elementary Geometry. V. 2: Stereometry, Transformation of Space [in Russian]. Moscow: MTsNMO, 2006.
2. *Volynets I. A. and Mednykh A. D.* Introduction to the Theory of Kleinian Groups [in Russian]. Omsk: OmGU, 1984.
3. *Fenchel W.* Elementary Geometry in Hyperbolic Space. Berlin; New York: Walter de Gruyter, 1989.
4. *McConnell B. D. S.* Hedronometric formulas for a hyperbolic tetrahedron // http://daylateanddollarshort.com/mathdocs/Hedronometric-Formulas-for-a-Hyperbolic-Tetrahedron.pdf.
5. *Abrosimov N. V. and Baigonakova G. A.* Hyperbolic octahedron with mmm-symmetry // Sib. Elektron. Mat. Izv. 2013. V. 10. P. 9–14.


*September 30, 2015*


E. S. Kudina,
Gorno-Altaĭsk State University, Gorno-Altaĭsk, Russia
eskudina@hotmail.com

A. D. Mednykh,
Sobolev Institute of Mathematics
Novosibirsk State University, Novosibirsk, Russia
mednykh@math.nsc.ru

Yakutian Mathematical Journal
October—December, 2015. Vol. 22, No. 4

UDC 514.16

# REGULAR POLYGONS AND POLYHEDRA OVER FINITE FIELDS

**T. M. Lavshuk**

**Abstract.** We establish necessary and sufficient conditions for the existence of regular polygons and polyhedra over finite fields of prescribed characteristic.

**Keywords:** regular polygon, regular polyhedron over finite field

## 1. Introduction

This article addresses the realization of regular polygons and regular polyhedra over the finite field $F_p$, the prime field of characteristic $p$. A regular polygon over $F_p$ is visually quite different from any ordinary one in the plane over $\mathbb{R}$. Only knowing the rules for constructing this object and conditions for its existence over $F_p$ enable us to chart further studies concerned with not just polygons. The problem of polygon construction is important in the theory of Riemann surfaces, to construct which we identify the edges of a fundamental polygon. To this end, we need to realize polygons in the Euclidean, hyperbolic, or spherical geometries. A similar problem arises in the three-dimensional case while constructing manifolds. A more general statement of the question is to study Riemann surfaces and manifolds over finite fields.

## 2. Regular Polygons over $F_p$ and Their Realizations

Define a regular polygon ($n$-gon) over $F_p$ using the concept of regular order $n$ star. Use Coxeter's method for constructing its vertices [1], that is, rotate an original vertex through the angle $\frac{2\pi}{n}$. In the case of a finite field replace the rotation by the reflection of points about a line.

Definition. If the tuple $[l_0, l_1, \dots, l_{n-1}]$ of lines is a regular order $n$ star with the point $O$ of common intersection then the set, obtained by this method, of $n$ points (which we call the vertices) with the condition that the squared distances between the adjacent points are congruent modulo $p$ and the lines (which we call the edges) passing through these points is called a *regular polygon* over the finite field $F_p$.

Let us determine the values of $p$ allowing realization of a regular triangle over $F_p$ using the theorem on the existence of order 3 star (see [2]).

**Theorem** (order 3 star). *A regular order 3 star exists only when 3 is a nonzero square.*

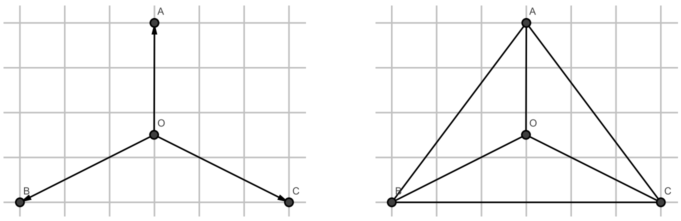
For a proof, see [2]. Let us establish, using a different approach, a necessary and sufficient condition for the existence of a regular triangle over $F_p$.

---

The author was supported by the Russian Foundation for Basic Research (Grant 15–01–07906).

**Theorem 2.1.** *A regular triangle over $F_p$ exists if and only if* 3 *is a nonzero square.*

PROOF. Consider three noncollinear vectors $\overrightarrow{OA}$, $\overrightarrow{OB}$, and $\overrightarrow{OC}$ in a plane as in Fig. 1($a$), where $O$ is the barycenter of the points $A$, $B$, and $C$. Construct a regular triangle over $\mathbb{R}$. Without loss of generality, assume that the coordinates of $O$ are $(0;0)$. The endpoints of the specified vectors and the lines passing through them form the regular triangle $ABC$ (Fig. 1($b$)).




($a$) ($b$)

Fig. 1. Construction of the regular triangle $ABC$

It is clear that

$$|\overrightarrow{OA}|^2 = |\overrightarrow{OB}|^2 = |\overrightarrow{OC}|^2, \quad \angle AOB = \angle BOC = \angle AOC.$$

This also holds over $F_p$ when we understand the equality of angles as the equality of the corresponding inner products. Consequently,

$$\overrightarrow{OA} \cdot \overrightarrow{OB} = \overrightarrow{OB} \cdot \overrightarrow{OC} = \overrightarrow{OA} \cdot \overrightarrow{OC}.$$

The above leads to the system

$$\begin{cases} (\overrightarrow{OA})^2 = (\overrightarrow{OB})^2 = (\overrightarrow{OC})^2, \\ \overrightarrow{OA} \cdot \overrightarrow{OB} = \overrightarrow{OB} \cdot \overrightarrow{OC} = \overrightarrow{OA} \cdot \overrightarrow{OC}, \\ \overrightarrow{OA} + \overrightarrow{OB} + \overrightarrow{OC} = 0. \end{cases} \tag{1}$$

It is equivalent to the system

$$\begin{cases} (\overrightarrow{OA})^2 = (\overrightarrow{OB})^2, \\ \overrightarrow{OA} \cdot \overrightarrow{OB} = -\frac{1}{2}(\overrightarrow{OA})^2. \end{cases} \tag{2}$$

If the vectors $\overrightarrow{OA}$ and $\overrightarrow{OB}$ have coordinates $(x, y)$ and $(s, t)$ then we can express (2) as

$$\begin{cases} x^2 + y^2 = s^2 + t^2, \\ xs + yt = -\frac{1}{2}(x^2 + y^2). \end{cases} \tag{3}$$

Solving (3), we express $x$ and $y$ in terms of $s$ and $t$:

$$\left[\begin{array}{l} \begin{cases} x = \frac{1}{2}(-s - \sqrt{3}t), \\ y = \frac{1}{2}(\sqrt{3}s - t), \end{cases} \\ \begin{cases} x = \frac{1}{2}(-s + \sqrt{3}t), \\ y = \frac{1}{2}(-\sqrt{3}s - t). \end{cases} \end{array}\right. \tag{4}$$

To realize a regular triangle $ABC$ over $F_p$, the coordinates of each of its vertices must exist. According to (4), for this condition to hold, we need $\sqrt{3}$ in $F_p$. We can always define $s$ and $t$ as elements of $F_p$. Consequently, we choose $p$ so that 3 is a nonzero square. Considering then that two noncollinear vectors constitute a basis for a plane, we conclude that for each suitable $p$ at least one regular triangle over $F_p$ exists.

We can make similar conclusions for regular pentagons over $F_p$ and regular heptagons over $F_p$ by the following theorem [2].

**Theorem** (order 5 star). *A regular order 5 star exists only when there exists a nonzero number $r$ satisfying the conditions* (i) $r^2 = 5$ *and* (ii) $2(5-r)$ *is a square.*

**Theorem** (order 7 star). *A regular order 7 star exists only when there exists a nonzero number $s$ such that* $7 - 56s + 112s^2 - 64s^3 = 0$ *and* $s(1-s)$ *is a square.*

It is difficult to recognize regular polygons over $F_p$ visually. For example, take a regular triangle over $F_p$ and inspect its construction.

Take the vertices $A\left(-\frac{1}{2}; \frac{\sqrt{3}}{2}\right)$, $B\left(-\frac{1}{2}; -\frac{\sqrt{3}}{2}\right)$, and $C(1;0)$ of a regular triangle over $\mathbb{R}$. According to Theorem 2.1 and the results of [3], choose as the characteristic $p$ the sequence $A038874$, that is, the primes congruent to $\{1, 2, 3, 11\}$ modulo 12. Assuming that $p = 11$ and recalculating the coordinates of the vertices of the specified triangle in $F_{11}$, we obtain $A(5;3)$, $B(5;8)$, and $C(1;0)$.

Now write down the equations of the lines passing through the adjacent vertices:

$$AB: \quad x = 5, \tag{5}$$

$$AC: \quad 8x + 4y + 3 = 0, \tag{6}$$

$$BC: \quad 3x + 4y + 8 = 0. \tag{7}$$

Find all points satisfying (5)–(7) and, together with $A$, $B$, and $C$, mark them in the lattice $F_{11}^2$.

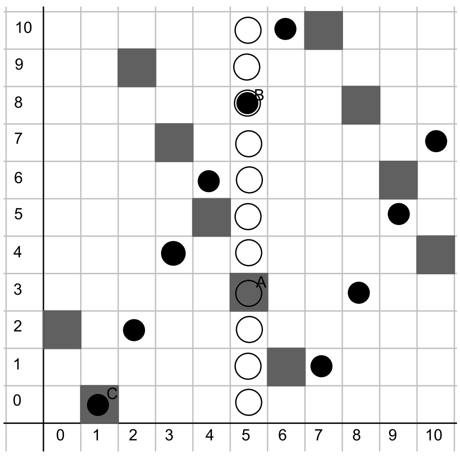



Fig. 2. A regular $\triangle ABC$ over $F_{11}$

The field $F_p$ lacks any order relation; therefore, for the lines determined by equations of the form $x = a$ or $y = b$ mark all lattice points satisfying them. For each line it is convenient to choose different labels (Fig. 2). According to the definition of regular $n$-gon over $F_p$ (in our case $n = 3$), we should verify that the squared side lengths of $\triangle ABC$ are congruent modulo 11. Using the definition of distance between points, we obtain

$$|AB|^2 \equiv |AC|^2 \equiv |BC|^2 \equiv 3 \pmod{11}.$$

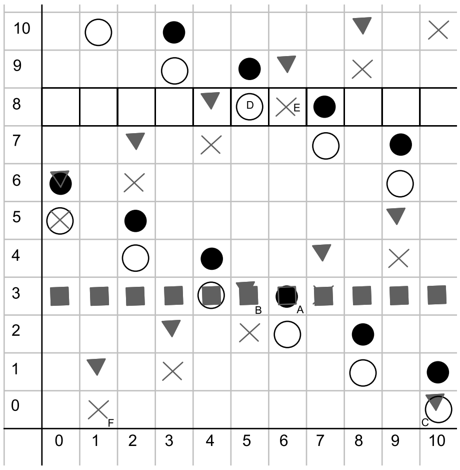



Fig. 3. A regular hexagon over $F_{11}$

Thus, a regular triangle over $F_{11}$ is realized.

Similarly we realize other regular polygons over $F_p$. Fig. 3 shows the result of this realization of a regular hexagon over $F_{11}$.

## 3. Regular Polyhedra in $F_p^3$ and Their Realizations

In this section we establish a criterion for the existence of regular polyhedra in $F_p^3$, which is the three-dimensional vector space over $F_p$. We also propose a visualization for the corresponding objects.

### 3.1. Regular tetrahedra in $F_p^3$.

Definition. Refer as a *regular tetrahedron* in $F_p^3$ to a set consisting of four regular triangles defined over $F_p$ and meeting pairwise along common edges.

Let us find out for which $p$ a regular tetrahedron in $F_p^3$ can be realized.

**Theorem 3.1.1.** *A regular tetrahedron is realized over every finite field $F_p$.*

Proof. Our goal is to show that for each prime $p$ at least one regular tetrahedron in $F_p^3$ exists.

Take three basis vectors $\overrightarrow{OA}$, $\overrightarrow{OB}$, and $\overrightarrow{OC}$ of length 1 in the three-dimensional space. To obtain a regular tetrahedron $ABCD$ over $\mathbb{R}$, joining the endpoints $A$, $B$, and $C$ firstly to each other by lines, we obtain a regular triangle $ABC$. Then join them to the point $D$ with coordinates $(t, t, t)$ as shown in Fig. 4.

Impose the equalities

$$|AB|^2 = |BC|^2 = |AC|^2 = |AD|^2 = |BD|^2 = |CD|^2.$$

The definition of distance between points yields

$$|AB|^2 = |BC|^2 = |AC|^2 = 2, \quad |AD|^2 = |BD|^2 = |CD|^2 = 3t^2 - 2t + 1.$$

Consequently,

$$3t^2 - 2t + 1 = 2.$$

Solving this equation, we obtain $t = -\frac{1}{3}$ or $t = 1$. Then $D$ has coordinates $(-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$ or $(1, 1, 1)$. To realize a regular tetrahedron in $F_p^3$, it is necessary

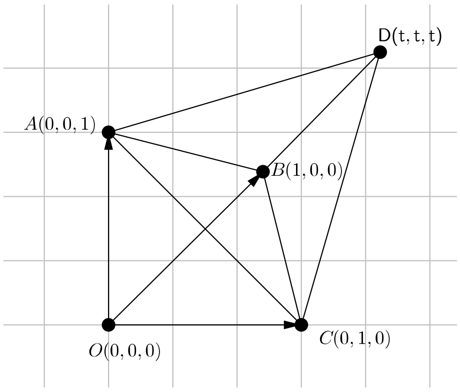



Fig. 4. A regular tetrahedron over $\mathbb{R}$

that the coordinates of all vertices exist. For $ABCD$ they exist in each field $F_p$. Therefore, for each $p$ there is always at least one regular tetrahedron realized in $F_p^3$, as required.

Realize in $F_p^3$ the regular tetrahedron with vertex coordinates

$$A\left(0; 0; \sqrt{\frac{2}{3}} - \frac{1}{2\sqrt{6}}\right), \quad B\left(-\frac{1}{2\sqrt{3}}; -\frac{1}{2}; -\frac{1}{2\sqrt{6}}\right),$$

$$C\left(-\frac{1}{2\sqrt{3}}; \frac{1}{2}; -\frac{1}{2\sqrt{6}}\right), \quad D\left(-\frac{1}{\sqrt{3}}; 0; -\frac{1}{2\sqrt{6}}\right).$$

As the characteristic $p$ we can choose the sequence of primes $A130063$ [3], that is, the primes congruent to 1 or 23 modulo 24. Assuming that $p = 23$ and calculating the vertex coordinates for a regular tetrahedron over $F_{23}$, we obtain $A(0; 0; 20)$, $B(5; 11; 1)$, $C(5; 12; 1)$, and $D(13; 0; 1)$.

Using the Wolfram Mathematica package, we construct this regular tetrahedron in $F_{23}^3$ (see Fig. 5).

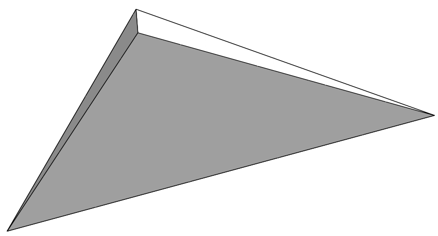


Fig. 5. A regular tetrahedron over $\mathbb{R}$

The squared distances between adjacent vertices of the constructed tetrahedron are congruent to 1 modulo 23. The vertices of each facet of the regular tetrahedron $ABCD$ in $F_{23}^3$ lie in one plane, which we check using the coplanarity criterion for three vectors. This applies to all its facets; furthermore, the planes of all facets are distinct. Thus, we have realized $ABCD$ in $F_{23}^3$.

### 3.2. Hexahedra in $F_p^3$.

DEFINITION. Refer as a *hexahedron* (*cube*) in $F_p^3$ to a set of points and lines forming six squares over $F_p$ of which each pair either meet along a common edge and two vertices or are disjoint; furthermore, each vertex belongs to three squares.

Let us prove a theorem on realization of a hexahedron in $F_p^3$, using which we establish a theorem on realization of other regular polyhedra in $F_p^3$.

**Theorem 3.2.1.** *A hexahedron is realized over every finite field $F_p$.*

PROOF. Taking three mutually orthogonal unit vectors in space, it is easy to construct a hexahedron over $\mathbb{R}$. The coordinates of their endpoints $A$, $B$, $C$, and $O$ are known (Fig. 6). The coordinates of the other points are $A_1(1;0;1)$, $B_1(1;1;1)$, $C_1(0;1;1)$, and $D(1;1;0)$.

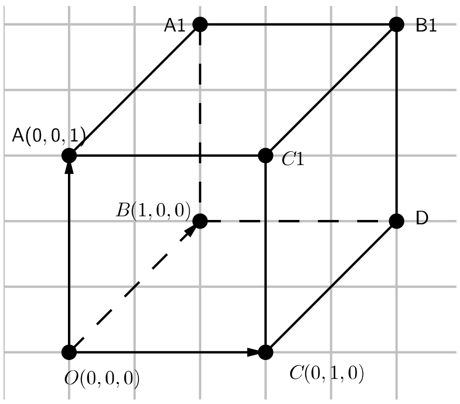



Fig. 6. A hexahedron over $\mathbb{R}$

Each of them exists over every finite field $F_p$. Consequently, for every $p$ we can always construct at least one hexahedron realized in $F_p^3$.

Observe that we can prove Theorem 3.1.1 using Theorem 3.2.1: inscribe a regular tetrahedron into a hexahedron. A hexahedron in $F_p^3$ looks as the usual cube over $\mathbb{R}$.

### 3.3. Octahedra in $F_p^3$.

DEFINITION. Refer as an *octahedron* in $F_p^3$ to a set of points and lines forming 8 regular triangles over $F_p$, each pair of which either meet along a common side and two vertices or are disjoint; furthermore, each vertex belongs to four triangles.

Using Theorem 3.2.1, we establish a similar result for octahedra.

**Theorem 3.3.1.** *An octahedron is realized over every finite field $F_p$ for $p > 2$.*

PROOF. Take the hexahedron $OBDCAA_1B_1C_1$ over $\mathbb{R}$ constructed in the proof of Theorem 3.2.1 (see Fig. 6). Inscribe in it an octahedron over $\mathbb{R}$. Its vertices are the midpoints of the facets of the hexahedron, which exist over every field $F_p$ for $p > 2$. Consequently, an octahedron is realized over every field $F_p$ for $p > 2$.

Fig. 7 shows a realization in $F_{17}^3$ of the octahedron with the vertices

$$A_1\left(-\frac{1}{\sqrt{2}};0;0\right),\quad A_2\left(0;\frac{1}{\sqrt{2}};0\right),\quad A_3\left(0;0;-\frac{1}{\sqrt{2}}\right),$$
$$A_4\left(0;0;\frac{1}{\sqrt{2}}\right),\quad A_5\left(0;-\frac{1}{\sqrt{2}};0\right),\quad A_6\left(\frac{1}{\sqrt{2}};0;0\right).$$

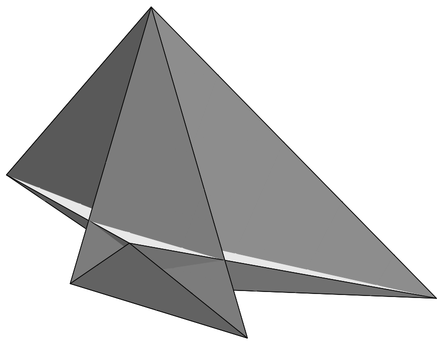


Fig. 7. An octahedron in $F_{17}^3$

### 3.4. Dodecahedra in $F_p^3$.

Definition. Refer as a *dodecahedron* in $F_p^3$ to the set of points and lines forming 12 regular pentagons over $F_p$, each pair of which either meet along a common edge and two vertices or are disjoint; furthermore, each vertex belongs to three pentagons.

**Theorem 3.4.1.** *A dodecahedron is realized over the finite field $F_p$ for $p \equiv \{0,1,4\} \pmod 5$.*

Proof. Take the hexahedron over $\mathbb{R}$ with the vertices

$$A_1\Big(\frac{1}{2};-\frac{1}{2};\frac{1}{2}\Big),\quad A_2\Big(\frac{1}{2};\frac{1}{2};\frac{1}{2}\Big),\quad A_3\Big(-\frac{1}{2};\frac{1}{2};\frac{1}{2}\Big),\quad A_4\Big(-\frac{1}{2};-\frac{1}{2};\frac{1}{2}\Big),$$

$$A_5\Big(\frac{1}{2};-\frac{1}{2};-\frac{1}{2}\Big),\quad A_6\Big(\frac{1}{2};\frac{1}{2};-\frac{1}{2}\Big),\quad A_7\Big(-\frac{1}{2};\frac{1}{2};-\frac{1}{2}\Big),\quad A_8\Big(-\frac{1}{2};-\frac{1}{2};-\frac{1}{2}\Big).$$

Inscribe it in a dodecahedron. Simple calculations yield the coordinates of its vertices different from the vertices of the hexahedron:

$$A_9\Big(\frac{\sqrt5-1}{4};0;\frac{\sqrt5+1}{4}\Big),\ A_{10}\Big(\frac{1-\sqrt5}{4};0;\frac{1+\sqrt5}{4}\Big),\ A_{11}\Big(0;\frac{\sqrt5+1}{4};\frac{\sqrt5-1}{4}\Big),$$

$$A_{12}\Big(0;\frac{\sqrt5+1}{4};\frac{1-\sqrt5}{4}\Big),\ A_{13}\Big(0;\frac{-1-\sqrt5}{4};\frac{\sqrt5-1}{4}\Big),\ A_{14}\Big(0;\frac{-1-\sqrt5}{4};\frac{1-\sqrt5}{4}\Big),$$

$$A_{15}\Big(\frac{1-\sqrt5}{4};0;\frac{-1-\sqrt5}{4}\Big),\ A_{16}\Big(\frac{\sqrt5-1}{4};0;\frac{-1-\sqrt5}{4}\Big),\ A_{17}\Big(\frac{-1-\sqrt5}{4};\frac{\sqrt5-1}{4};0\Big),$$

$$A_{18}\Big(\frac{-1-\sqrt5}{4};\frac{1-\sqrt5}{4};0\Big),\ A_{19}\Big(\frac{1+\sqrt5}{4};\frac{\sqrt5-1}{4};0\Big),\ A_{20}\Big(\frac{1+\sqrt5}{4};\frac{1-\sqrt5}{4};0\Big).$$

To realize a dodecahedron in $F_p^3$, we need the coordinates of all its vertices to exist in the specified field. The existence of coordinates of the points coinciding with the vertices of the hexahedron follows from Theorem 3.2.1.

For the existence of the remaining vertices, whose coordinates are obtained above, we need 5 to be a square modulo $p$. Consequently, considering that every three noncoplanar vectors constitute a basis for space, for each $p$ defined by the sequence *A*038872 [3], that is, for each prime congruent to $\{0,1,4\}$ modulo 5, there is always at least one dodecahedron which is realized in $F_p^3$.

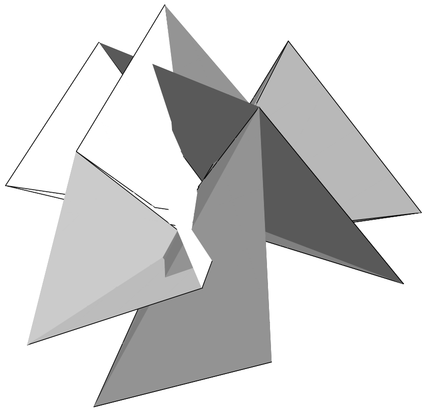


Fig. 8. A dodecahedron in $F_{11}^3$

Fig. 8 shows the realization in $F_{11}^3$ of the dodecahedron $A_1A_2\dots,A_{20}$ indicated in the proof of Theorem 3.4.1.

### 3.5. Icosahedra in $F_p^3$.

Definition. Refer as an *icosahedron* in $F_p^3$ to a set of points and lines forming 20 regular triangles over $F_p$, each pair of which either meet along a common edge and two vertices or have only one common vertex or are disjoint; furthermore, each vertex belongs to five triangles.

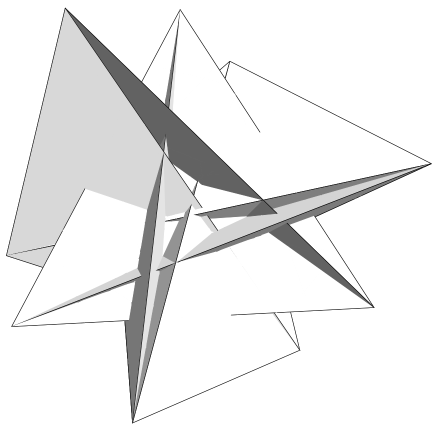


Fig. 9. An icosahedron in $F_{11}^3$

The icosahedron and dodecahedron are dual polyhedra. We can inscribe an icosahedron into a dodecahedron so that then the vertices of the icosahedron are the midpoints of the facets of the dodecahedron. Then, excluding the case $p = 5$, we can justify the following statement.

**Theorem 3.5.1.** *An icosahedron is realized over the finite field $F_p$ for $p > 5$ with $p \equiv \{0, 1, 4\}$ (mod 5).*

To construct an icosahedron in $F_p^3$, take the icosahedron $A_1A_2\cdots A_{12}$ over the field $\mathbb{R}$ with vertex coordinates

$$A_1\left(\frac{5+3\sqrt{5}}{20};0;\frac{-5-\sqrt{5}}{20}\right),\ A_2\left(\frac{-5-3\sqrt{5}}{20};0;\frac{5+\sqrt{5}}{20}\right),\ A_3\left(\frac{5+\sqrt{5}}{20};\frac{-5-3\sqrt{5}}{20};0\right),$$

$$A_4\left(\frac{-5-\sqrt{5}}{20};\frac{5+3\sqrt{5}}{20};0\right),\ A_5\left(0;\frac{5+\sqrt{5}}{20};\frac{-5-3\sqrt{5}}{20}\right),\ A_6\left(\frac{5+\sqrt{5}}{20};\frac{5+3\sqrt{5}}{20};0\right),$$

$$A_7\left(\frac{-5-\sqrt{5}}{20};\frac{-5-3\sqrt{5}}{20};0\right),\ A_8\left(0;\frac{-5-\sqrt{5}}{20};\frac{5+3\sqrt{5}}{20}\right),\ A_9\left(0;\frac{-5-\sqrt{5}}{20};\frac{-5-3\sqrt{5}}{20}\right),$$

$$A_{10}\left(\frac{5+3\sqrt{5}}{20};0;\frac{5+\sqrt{5}}{20}\right),\ A_{11}\left(\frac{-5-3\sqrt{5}}{20};0;\frac{-5-\sqrt{5}}{20}\right),\ A_{12}\left(0;\frac{5+\sqrt{5}}{20};\frac{5+3\sqrt{5}}{20}\right).$$

As the characteristic $p$ we can choose the sequence $A038872$ [3].

Assuming that $p = 11$ and calculating the vertex coordinates of an icosahedron over $F_{11}$, we obtain $A_1(8;0;10)$, $A_2(3;0;1)$, $A_3(1;3;0)$, $A_4(10;8;0)$, $A_5(0;1;3)$, $A_6(1;8;0)$, $A_7(10;3;0)$, $A_8(0;10;8)$, $A_9(0;10;3)$, $A_{10}(8;0;1)$, $A_{11}(3;0;10)$, $A_{12}(0;1;8)$.

Using the Wolfram Mathematica package, we construct this icosahedron in $F_{11}^3$ (see Fig. 9).

We expect to obtain similar results for semiregular polyhedra.

## *REFERENCES*


1. *Coxeter H. S. M.* Introduction to Geometry. New York etc.: John Wiley and Sons, 1961.
2. *Wildberger N. J.* Divine Proportions: Rational Trigonometry to Universal Geometry. Australia: Wild Egg Pty Ltd., 2005.
3. *Sloane N. J. A.* The On-Line Encyclopedia of Integers Sequences. 1964.


*October 15, 2015*

T. M. Lavshuk
Novosibirsk State University, Novosibirsk, Russia
`tamara_090391@mail.ru`

*Yakutian Mathematical Journal*
*October—December, 2015. Vol. 22, No. 4*

*UDC 514.123*

# THE EXISTENCE OF A EUCLIDEAN STRUCTURE ON THE FIGURE–EIGHT KNOT WITH A BRIDGE

**A. D. Mednykh and D. Yu. Sokolova**

**Abstract.** This paper studies the main geometric invariants of the Euclidean conical manifold whose singular set is the figure-eight knot with a bridge and whose support is the three-dimensional sphere. We obtain conditions for the existence of this manifold and calculate its volume and the lengths of its singular geodesics.

**Keywords:** cone-manifold, Euclidean structure, volume, figure-eight knot

## 1. Introduction

Koebe's classical articles [1] established uniformization theorems for the Riemannian surfaces of a prescribed signature. In the modern language this means that all "good" two-dimensional orbifolds have universal coverings amounting to either the unit disk or the complex plane or the Riemann sphere. From the viewpoint of geometry, the latter is equivalent to the assertion that every "good" two-dimensional orbifold carries either hyperbolic or Euclidean or spherical structure. Recall that an orbifold is "good" whenever it is neither a sphere with one singular point nor a sphere with two singular points of distinct orders. The analog in the three-dimensional case, more complicated to state, is called Thurston's geometrization conjecture. It is completely settled by the Russian mathematician Perelman [2–4]; as a particular case, this yielded a solution to the famous Poincaré conjecture.

We can express manifolds and orbifolds with a geometric structure as quotients spaces $X/\Gamma$, where $X$ is one of the known geometries and $\Gamma$ is a discrete group of isometries acting on $X$ with fixed points in general. In low dimensions all possible geometries are known. In particular, in the two-dimensional case $X = \mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2$, and in the three-dimensional case it is one of eight model geometries of Thurston: $X = \mathbb{S}^3, \mathbb{E}^3, \mathbb{H}^3, \mathbb{S}^2 + \mathbb{E}^1, \mathbb{H}^2 + \mathbb{E}^1, \mathscr{N}il, \mathscr{S}ol, \widetilde{\mathscr{P}\mathscr{S}\mathscr{L}}(2, \mathbb{R})$.

Assume that $X$ is one of the listed three-dimensional geometries. Then the image of the fixed points of $\Gamma$ under the canonical mapping $X \mapsto X/\Gamma$ is typically a knot, a link, or a knotted graph. Let us illustrate this with one example [5]. Take $X = \mathbb{H}^3$ and $\Gamma = \mathbb{F}_{2n}$, where $n \geq 4$, that is, the Fibonacci group acting on $X$ by isometries. Then $X/\Gamma$ is the three-dimensional sphere, while the image of the fixed-point set of $X$ in $X/\Gamma$ is a figure-eight knot.

But in general the presence of a geometric structure is not necessarily related to discrete groups. This leads to conical manifolds, which we can regard as straightforward generalizations of orbifolds. In turn, in the definition of conical manifold in the

The authors were supported by the Russian Foundation for Basic Research; the first, by Grant 15–01–07906 and the second, by Grant 16–31–00138.

following section we only need local uniformization using the geometries mentioned above.

The goal of this article is to study Euclidean structures on knots and links. In 1975 Riley discovered [6] examples of hyperbolic structures on some knots and the complements to some links in the three-dimensional sphere. In the spring of 1977 Thurston presented an existence theorem for a Riemannian metric of constant negative curvature on a three-dimensional manifold. In practice it turned out that the complement to each prime knot excluding torus knots and satellite knots admits a hyperbolic structure. Note the following available results. A Euclidean structure on the figure-eight knot $4_1$ arises when its conical angle $\alpha$ equals $\frac{2\pi}{3}$; this is due to Thurston [7]. Mednykh and Rasskazov gave in [8] an explicit construction of the fundamental set for the conical manifold $4_1(\alpha)$ in $E^3$. This fundamental set amounts to a nonconvex polyhedron with 20 sides whose vertices have integer coordinates. Shmatkov studied in [9] the existence of Euclidean structure on the Whitehead link. The structure of fundamental polyhedron for the trefoil knot with a bridge was worked out in [10] and conditions for the existence of Euclidean structure on the corresponding cone-manifold were found.

In this article we study the main geometric invariants of the Euclidean cone-manifold whose singular set is the figure-eight knot with a bridge, and the support is the three-dimensional sphere. We establish conditions for the existence of this manifold and calculate its volume and the lengths of its singular geodesics.

## 2. Preliminaries

A *three-dimensional cone-manifold* is a metric space obtained from a collection of disjoint 3-simplices in a space of constant sectional curvature $k$ by an isometric identification of their facets. Furthermore, we assume that the resulting topological space (the support) is a manifold.

This manifold is equipped with a Riemannian metric of constant sectional curvature $k$ on the union of cells of dimension 2 and 3. In the case $k = 0$ say that the corresponding cone-manifold *has* (or *admits*) *Euclidean structure*. Similarly define cone-manifolds with spherical ($k = +1$) and hyperbolic ($k = -1$) structures.

The metric structure near each 1-cell is determined by the *conical angle*, which is the sum of dihedral angles for the edges whose identification produces this cell.

Refer as the *singular set* of a cone-manifold to the closure of all 1-cells the conical angle around which differs from $2\pi$.

We should also note that a point in the singular set with conical angle $\alpha$ has a neighborhood isometric to a neighborhood of a point lying on the edge of a wedge with opening angle $\alpha$ whose sides are pairwise identified by way of rotating the three-dimensional space about the edge of the wedge. We can visualize a conical manifold as a three-dimensional manifold with an embedded graph on which the metric is distorted. Furthermore, if we measure the length of an infinitesimal circle around a component of the graph then instead of the standard $2\pi\varepsilon$ we obtain $\alpha\varepsilon$, where $\alpha$ is the conical angle along the component of the graph.

Let us define the *holonomy group* of a geometric orbifold. Take a geometric orbifold $\mathscr{O}$ possessing a $(G, X)$-structure [7]. Consider the associated $(G, X)$-manifold $M = \mathscr{O} \setminus \Sigma$, where $\Sigma$ is the singular set of $\mathscr{O}$. Take regions $U_1, U_2, \dots$ and mappings $\varphi_i : U_i \to X$ determining local coordinate systems on $M$ with transition functions

$$\gamma_{ij} = \varphi_i \circ \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j).$$

By the definition of $(G, X)$-manifold, each mapping $\gamma_{ij}$ acts locally as an element of $G$, so that we can consider $\gamma_{ij}$ as a locally constant mapping with values in $G$. The composition with $\varphi_j$ yields a locally constant mapping $U_i \cap U_j \to G$, which we also denote by $\gamma_{ij}$.

Suppose now that two charts $(U_i, \varphi_i)$ and $(U_j, \varphi_j)$ cover the same point $x$. Then we can modify the mapping $\varphi_j$ (by considering its composition with $\gamma_{ij}$) so that it coincides with the mapping $\varphi_i$ near $x$. Actually, if the intersection $U_i \cap U_j$ is connected then these mappings coincide on the whole intersection, so that we obtain a mapping $U_i \cap U_j \to X$ extending $\varphi_i$. But, in general, attempting in this fashion to extend the coordinate mapping to the entire manifold, we arrive at mismatching values. In order to avoid that, we must pass to the universal covering space.

Choose a basepoint $x_0 \in M$ and a chart $(U_0, \varphi_0)$ covering it. Take the universal covering space $\pi : \widetilde{M} \to M$ of $M$. Regard $\widetilde{M}$ as the space of homotopy classes of paths in $M$ beginning at $x_0$ and consider a path $\alpha$ representing a homotopy class $[\alpha] \in \widetilde{M}$ (so that $\alpha(1) = \pi([\alpha])$). Subdivide $\alpha$ by the intermediate points

$$x_0 = \alpha(t_0),\ x_1 = \alpha(t_1),\ \dots,\ x_n = \alpha(t_n),$$

where $t_0 = 0$ and $t_1 = 1$, so that each of the fragments is covered by some chart $(U_i, \varphi_i)$. Then, moving along $\alpha$, we modify the next mapping $\varphi_i$ to coincide with the (already modified) mapping $\varphi_{i-1}$ in some neighborhood of $x_i \in U_{i-1} \cap U_i$. Agreeing with each another, these charts constitute the *analytic continuation* of the mapping $\varphi_0$ along this path. The last of the new coordinate mappings is of the form

$$\psi = \gamma_{01}(x_1)\gamma_{12}(x_2)\dots\gamma_{n-1,n}(x_n)\varphi_n.$$

Fixing a base point and an initial mapping $\varphi_n$, define the *development mapping* $D : \widetilde{M} \to X$ as the mapping locally specified using the analytic continuation of $\varphi_0$ along each path; that is, $D = \varphi_0^\sigma \circ \pi$ in some neighborhood of $\sigma \in \widetilde{M}$.

When initial conditions (basepoint and initial mapping) change, the image of the development mapping changes under the action of some element of $G$.

If we endow the space with the universal covering $(G, X)$-structure induced by the covering $\pi$ then the development mapping is a local $(G, X)$-homeomorphism between $\widetilde{M}$ and $X$.

Although in the most interesting cases the group $G$ acts on $X$ transitively, this condition is not necessary for the definition of $D$. For instance, if the group $G$ is trivial and the manifold $X$ is closed then the closed $(G, X)$-manifold is precisely a finite covering of $X$ with projection $D$.

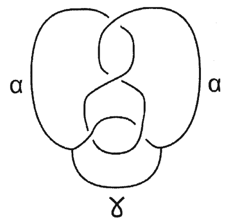



Fig. 1. Figure-eight knot with a bridge

Consider now an element $\sigma$ of the fundamental group of $M$. Analytic continuation along the loop $\sigma$ leads to the germ $\varphi_0^\sigma$, which we can compare with $\varphi_0$ since they are both defined in a neighborhood of the basepoint. Denote by $g_\sigma$ an element of $G$ with $\varphi_0^\sigma = g_\sigma \varphi_0$ and call $g_\sigma$ the *holonomy* of $\sigma$. It is easy to deduce from the definition of development mapping that $D \circ T_\sigma = g_\sigma \circ D$, where $T_\sigma : \tau \to \sigma\tau$ is the transformation of the covering induced by $\sigma$. Applying this equality to a product of loops, we infer that the mapping $H : \sigma \to g_\sigma$ from $\pi_1(M)$ to $G$ is a homomorphism, which we call the *holonomy* of $M$. Its image is called the *holonomy group* of $M$. Note that $H$ depends on the choice in constructing $D$: when $D$ changes, the image of $H$ is conjugated by an element of $G$.

This article studies the cone-manifold $\mathscr{O} = \mathscr{O}(\alpha, \alpha; \gamma)$ whose support is the three-dimensional sphere $\mathbb{S}^3$ and the singular set $\Sigma$ is the figure-eight knot with one bridge, which amounts to the graph in Fig. 1.

We can find the fundamental group $\pi_1(\mathbb{S}^3 \setminus \Sigma)$ of the complement to the graph using Wirtinger's algorithm. It has two generators. We study the geometric structure on this cone-manifold.

The cone-manifold is a completion of a metric space with incomplete Euclidean metric. The value of conical angle $\alpha$ along the knot component is determined by the completion of the metric space. The latter means that if homeomorphisms $g$ and $h$ carry a neighborhood of a point of a manifold into balls of the form $B^3 = \{x \in \mathbb{R}^3 : \|x\| < 1\}$ with the surface area element $ds^2 = dx^2 + dy^2 + dz^2$ then the homeomorphism $g \circ h^{-1}$ consists of motions of the Euclidean space. Therefore, it preserves the Euclidean metric. Furthermore, representing the generators of the fundamental group by rotation matrices in the Euclidean space, we obtain conditions for the existence of Euclidean structure on the cone-manifold. To this end, find the holonomy group of this manifold.

Consider the holonomy mapping $\varphi : \pi_1(\mathbb{S}^3 \setminus \Sigma) \to \mathrm{Isom}(\mathbb{E}^3)$ carrying the generators $s$ and $t$ of the fundamental group

$$\pi_1(\mathbb{S}^3 \setminus \Sigma) = \langle s, t : s\ell_s = \ell_s s\rangle, \quad \text{where } \ell_s = stst^{-1}s^{-1}tsts^{-1}t^{-1},$$

of the knot to the linear transformations

$$\mathscr{S}(x) = (x - e_3)S + e_3, \quad \mathscr{T}(x) = (x + e_3)T - e_3 \tag{1}$$

respectively, where $e_3 = (0, 0, 1)$, while $S$ and $T$ are rotation matrices.

Following [9], put $M = \cot \frac{\alpha}{2}$. Then $S$ and $T$ become

$$S = \frac{1}{M^2+1}\begin{pmatrix} M^2 + \cos\theta & \sin\theta & -2M\sin\frac{\theta}{2} \\ \sin\theta & M^2 - \cos\theta & 2M\cos\frac{\theta}{2} \\ 2M\sin\frac{\theta}{2} & -2M\cos\frac{\theta}{2} & -1 + M^2 \end{pmatrix}, \tag{2}$$

$$T = \frac{1}{M^2+1}\begin{pmatrix} M^2 + \cos\theta & -\sin\theta & -2M\sin\frac{\theta}{2} \\ -\sin\theta & M^2 - \cos\theta & -2M\cos\frac{\theta}{2} \\ 2M\sin\frac{\theta}{2} & 2M\cos\frac{\theta}{2} & -1 + M^2 \end{pmatrix}, \tag{3}$$

where $\theta$ is the angle of relative rotation between the singular components.

Assume furthermore that the holonomy mapping carries the element $\ell_s$ into the rotation through angle $\gamma$ about the singular component corresponding to the bridge of the knot.

Refer as the *holonomy group* of the manifold under study to the group generated by the rotations $\mathscr{S}$ and $\mathscr{T}$ through angle $\alpha$ about the singular component of the fundamental set.

## 3. Structure of the Fundamental Set for the Figure-Eight Knot with a Bridge

Let us construct some fundamental set for the manifold $\mathscr{O}(\alpha, \alpha; \gamma)$. Consider the collection of disjoint 3-simplices in the space of constant zero curvature shown in Fig. 2, from which the cone-manifold results by way of isometric identification of facets. The fundamental set is the polyhedron $\mathscr{F}$, with 12 vertices and 20 facets, which we obtain by gluing simplices along common edges $Q_0Q_1$.

This set can be realized in all geometries: $\mathbb{S}^3$, $\mathbb{H}^3$, and $\mathbb{E}^3$. Identify the curvilinear facets of $\mathscr{F}$ via isometric transformations $\mathscr{S}$ and $\mathscr{T}$ using the rule

$$\mathscr{S} : P_1P_0P_9P_8P_7P_6 \to P_1P_2P_3P_4P_5P_6, \quad \mathscr{T} : P_4P_5P_6P_7P_8P_9 \to P_4P_3P_2P_1P_0P_9.$$

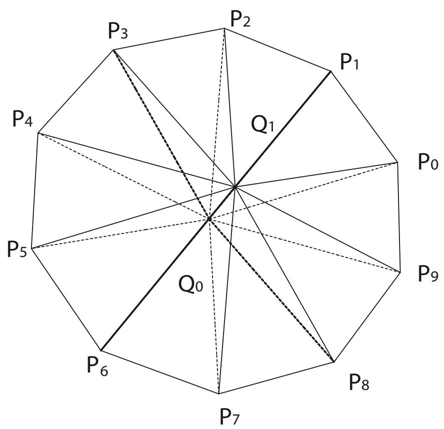



Fig. 2. The fundamental icosahedron $\mathscr{F}$

## 4. Realization of the Fundamental Set in Euclidean Space

Let us describe some geometric realization of the fundamental set $\mathscr{O}(\alpha, \alpha; \gamma)$ in the Euclidean space. To this end, express the coordinates of its vertices in terms of certain parameters possessing geometric meaning.

Put $X = \cos\frac{\theta}{2}$ and $Y = \sin\frac{\theta}{2}$, where $\theta$ is the angle of relative rotation between the knot components. Then the fixed-point sets of $\mathscr{S}$ and $\mathscr{T}$ in (1) are the lines

$$\mathrm{Fix}(\mathscr{S}) = (tX, tY, 1), \quad \mathrm{Fix}(\mathscr{T}) = (tX, -tY, -1), \quad t \in \mathbb{R}.$$

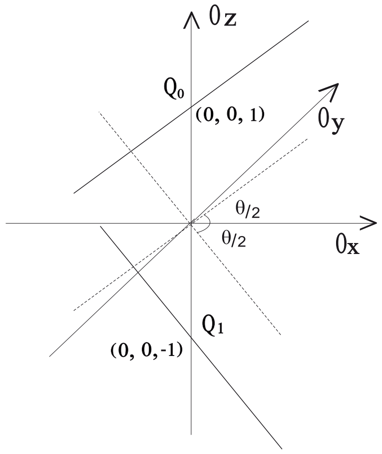



Fig. 3. The rotational axes $\mathrm{Fix}(\mathscr{S})$ and $\mathrm{Fix}(\mathscr{T})$

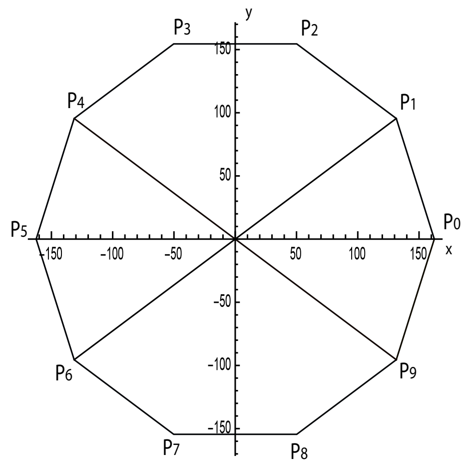



Fig. 4. The projection of $\mathscr{F}$ to the $Oxy$ plane

In the three-dimensional Euclidean space the rotational axes $\mathrm{Fix}(\mathscr{S})$ and $\mathrm{Fix}(\mathscr{T})$ are skew lines perpendicular to the $Oz$ axis and with angle $\theta$ between them (Fig. 3).

For the fundamental polyhedron $\mathscr{F}$ (see Fig. 2) the pairs of vertices $P_1, P_6$ and $P_4, P_9$ lie respectively on the axes $\mathrm{Fix}(\mathscr{S})$ and $\mathrm{Fix}(\mathscr{T})$.

The figure-eight knot with a bridge has three second-order symmetries. On the fundamental polyhedron they are realized as rotations about the $Ox$, $Oy$, and $Oz$ axes. In particular, the second-order rotation about the $Ox$ axis leaves the fundamental polyhedron invariant. This implies that two vertices of $\mathscr{F}$ lie on the $Ox$ axis.

According to Fig. 4, we can represent the coordinates of the vertices of $\mathscr{F}$ as follows:

$$\begin{gathered} P_0 = (x, 0, 0), \quad P_1 = (tX, tY, 1), \quad P_2 = (a, b, c), \\ P_3 = (-a, b, -c), \quad P_4 = (-tX, tY, -1), \quad P_5 = (-x, 0, 0), \\ P_6 = (-tX, -tY, 1), \quad P_7 = (-a, -b, c), \quad P_8 = (a, -b, -c), \\ P_9 = (tX, -tY, -1), \quad Q_0 = (0, 0, 1), \quad Q_1 = (0, 0, -1). \end{gathered} \tag{4}$$

Observe that $P_2 = P_0\mathscr{S} = P_6\mathscr{T}$. Rearrange this as

$$\begin{cases} (a, b, c) = (x, 0, 0)\mathscr{S}, \\ (x, 0, 0)\mathscr{S} = (-tX, -tY, 1)\mathscr{T}. \end{cases} \tag{5}$$

Solving the second equation of this system for $x$ and $t$ and recalling that $X^2+Y^2 = 1$, we obtain

$$x = \frac{5 + 4M^2 - M^4 - 20X^2 - 4M^2X^2}{2MY(1 + M^2 - 8X^2)}, \quad t = \frac{X(3M^2 - 5)}{MY(1 + M^2 - 8X^2)}. \tag{6}$$

Furthermore, comparing the first coordinates of $(x, 0, 0)\mathscr{S}$ and $(-tX, -tY, 1)\mathscr{T}$ and using again the equality $X^2 + Y^2 = 1$, we infer that $M$ and $X$ satisfy

$$5 + 6M^2 + M^4 - 60X^2 - 12M^2X^2 + 80X^4 = 0. \tag{7}$$

Inserting (6) and (7) into the first equation of (5), we find the values of $a$, $b$, and $c$:

$$\begin{gathered} a = \frac{4M^2 - 15X^2 - 7M^2X^2 + 20X^4}{MY(1 + M^2 - 8X^2)}, \\ b = \frac{X(5 + M^2 - 20X^2)}{M(1 + M^2 - 8X^2)}, \quad c = \frac{M^2 + 4X^2 - 3}{1 + M^2 - 8X^2}. \end{gathered} \tag{8}$$

## 5. The Euclidean Volume of the Cone-Manifold

The following two theorems are the main results of this article.

**Theorem 1.** *If $\alpha \in [\frac{2\pi}{3}, \pi)$ then for some $\gamma \in (0, 2\pi]$ the cone-manifold $\mathscr{O}(\alpha, \alpha; \gamma)$ carries a Euclidean structure. In particular, $\mathscr{O}(\frac{2\pi}{3}, \frac{2\pi}{3}, 2\pi) = 4_1(\frac{2\pi}{3})$ is a Euclidean orbifold whose support is a three-dimensional sphere, and whose singular set is the figure-eight knot with conical angle $\frac{2\pi}{3}$.*

Proof. It suffices to establish that for each $\alpha \in [\frac{2\pi}{3}, \pi)$ there exists a polyhedron $\mathscr{F}$, described in the previous section, in which the sum of dihedral angles along the inscribed edges $P_iP_{i+1}$, for $i = 0, \dots, 9$, equals $\gamma \in (0, 2\pi]$. To this end, put $M = \cot\frac{\alpha}{2}$ and $X = \cos\frac{\theta}{2}$ and consider the curve in Fig. 5 defined by (5).

As the main relation between $\alpha$ and $\theta$, the equation (7) of the curve is obtained from (5). Moreover, as becomes clear below, its part highlighted in Fig. 5 corresponds to the modeled situation.

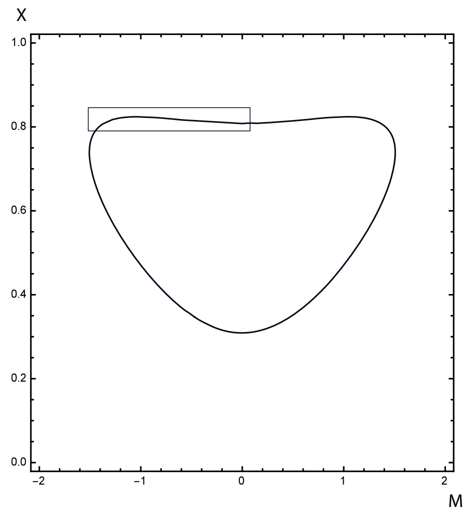



Fig. 5. Curve of the existence of a Euclidean structure for $\mathscr{O}(\alpha,\alpha;\gamma)$

**Lemma 1.** *Suppose that $M \in \left(0, \frac{1}{\sqrt{3}}\right]$ and $X \in \left(\sqrt{\frac{3+\sqrt{5}}{8}}, \sqrt{\frac{2}{3}}\right]$ and that the main relation (7) holds. Then there exists a fundamental polyhedron $\mathscr{F}$ (Fig. 2) determined by the parameters $\alpha$ and $\theta$, where $M = \cot\frac{\alpha}{2}$ and $X = \cos\frac{\theta}{2}$.*

Proof. First of all, verify that this polyhedron exists for $\alpha = \frac{2\pi}{3}$ and $\theta = 2\cos^{-1}\sqrt{\frac{2}{3}}$. Indeed, in this case the vertices have integer coordinates equal to

$$P_0 = (3,0,0),\quad P_1 = (2,\sqrt{2},1),\quad P_2 = (1,\sqrt{8},0),\quad P_3 = (-1,\sqrt{8},0),$$
$$P_4 = (-2,\sqrt{2},-1),\quad P_5 = (-3,0,0),\quad P_6 = (-2,-\sqrt{2},1),$$
$$P_7 = (-1,-2\sqrt{2},0),\quad P_8 = (1,-\sqrt{8},0),\quad P_9 = (2,-\sqrt{2},-1).$$

Furthermore, $\mathscr{F}$ is fundamental for the orbifold $4_1(\frac{2\pi}{3})$ with support $S^3$ and singular set equal to a figure-eight knot with conical angle $\frac{2\pi}{3}$. Its structure is described in detail in [8]. In particular, the oriented volumes $V_i$ of the tetrahedra $Q_0Q_1P_iP_{i+1}$ for $i = 0,\dots,9$ are positive, and the interiors of these polyhedra are disjoint.

Recall that the oriented volume of the tetrahedron $T$ with vertices $(x_j, y_j, z_j)$, for $j = 1,2,3,4$, satisfies

$$\operatorname{Vol} T = \frac{1}{6}\det\begin{pmatrix} x_1 - x_4 & y_1 - y_4 & z_1 - z_4 \\ x_2 - x_4 & y_2 - y_4 & z_2 - z_4 \\ x_3 - x_4 & y_3 - y_4 & z_3 - z_4 \end{pmatrix}.$$

We can extract the following three forms of the formula for the volume $V_i$:

$$\begin{aligned} V_0 = V_9 = V_4 = V_5 &= \frac{txY}{3}, \\ V_2 = V_7 &= \frac{2}{3}ab, \\ V_1 = V_3 = V_6 = V_8 &= \frac{1}{3}t(Xb - Ya). \end{aligned} \tag{9}$$

Consider the regions of the degeneration of $V_i$ using (6), (7) and (8).

For $M \in (0, \frac{1}{\sqrt{3}}]$ and $X \in (\sqrt{\frac{3+\sqrt{5}}{8}}, \sqrt{\frac{2}{3}}]$ the oriented volumes $V_i$ keep the signs and remain positive; consequently, under the conditions of the lemma they cannot degenerate. Therefore, the condition that $V_i > 0$ for $i = 0, \dots, 9$ is equivalent to the inequality $a > 0$, where $a$ is given by (8).

**Corollary.** *If the manifold $\mathscr{O}(\alpha, \alpha; \gamma)$ is Euclidean then*

$$\cos\gamma = \frac{1}{1953125}\Bigg(\frac{1}{(1+M^2)^{10}} 128M^2(M^2+5)^2(11M^2-25)$$
$$\times(3125 - 21875M^2 + 1250M^4 - 9750M^6 - 11175M^8 - 2823M^{10})X^2$$
$$-\frac{169869312}{(1+M^2)^9} + \frac{254803968}{(1+M^2)^8} + \frac{23461888}{(1+M^2)^7} - \frac{136282112}{(1+M^2)^6} - \frac{10575872}{(1+M^2)^5}$$
$$+\frac{56000512}{(1+M^2)^4} + \frac{2232832}{(1+M^2)^3} - \frac{14626688}{(1+M^2)^2} - \frac{4716288}{(1+M^2)} + 1524197\Bigg).$$

Proof. Consider the commutator $K = STST^{-1}S^{-1}TSTS^{-1}T^{-1}$ corresponding to the word $\ell_s = stst^{-1}s^{-1}tsts^{-1}t^{-1}$. The matrix $K$ amounts to the rotation through angle $\gamma$ about some edge $P_iP_{i+1}$ corresponding to the bridge between the components of the figure-eight knot (see Fig. 1). The trace of the orthogonal matrix $K$ is related to the rotation angle as $\operatorname{tr} K = 2\cos\gamma + 1$. Simplifying the expression $\cos\gamma = \frac{1}{2}(\operatorname{tr} K - 1)$, we obtain the original equality.

**Theorem 2.** *The Euclidean volume of the cone-manifold $\mathscr{O}(\alpha, \alpha; \gamma)$ equals*

$$\operatorname{Vol}(\mathscr{O}(\alpha, \alpha; \gamma)) = \frac{8X\sqrt{1-X^2}(M^4 - 50M^2X^2 + 150X^2 - 25)}{3M^2(1+M^2-8X^2)^2}.$$

Proof. The Euclidean volume of the cone-manifold $\mathscr{O}(\alpha, \alpha; \gamma)$ equals the volume of the fundamental polyhedron $\mathscr{F}$ depicted in Fig. 2. Therefore, the Euclidean volume $\operatorname{Vol}(\mathscr{O}(\alpha, \alpha; \gamma))$ amounts to the sum of the volumes $V_i$ of the tetrahedra $Q_0Q_1P_iP_{i+1}$, where $i = 0, \dots, 9$, and $P_{10} = P_0$, and we can find it using (6), (8) and (9):

$$\operatorname{Vol}(\mathscr{O}(\alpha, \alpha; \gamma)) = \sum_{i=0}^{9} V_i = \frac{4}{3}(ab + t(Y(x-a) + Xb))$$
$$= -\frac{16X\sqrt{1-X^2}(15 + 3M^2 - 105X^2 + 19M^2X^2 + 40X^4)}{3M^2(1+M^2-8X^2)^2}.$$

The remainder of the division of $P = 15 + 3M^2 - 105X^2 + 19M^2X^2 + 40X^4$ by $Q = 5 + 6M^2 + M^4 - 60X^2 - 12M^2X^2 + 80X^4$ equals $M^4 - 50M^2X^2 + 150X^2 - 25$. Since in our case $Q = Q(M, X) \equiv 0$, we can express the answer as

$$\operatorname{Vol}(\mathscr{O}(\alpha, \alpha; \gamma)) = \frac{8X\sqrt{1-X^2}(M^4 - 50M^2X^2 + 150X^2 - 25)}{3M^2(1+M^2-8X^2)^2}.$$

Let us illustrate the results with an example.

## 6. Example

The table below shows the results of numerical experiments. We also calculated:

—the lengths of singular geodesics $\ell_\alpha$ and $\ell_\gamma$, equal respectively to $\ell_\alpha = 2t$, where $t$ is given in (6), and $\ell_\gamma = \sum_{i=0}^{9} |P_iP_{i+1}|$.

**Table 1**

| conical angle $\alpha$ of the manifold $\mathscr{O} = \mathscr{O}(\alpha, \alpha, \gamma)$, parameters $X = \cos\frac{\theta}{2}$, $g = \cos\gamma$ | Euclidean volume $\mathrm{Vol}(\mathscr{O})$ and reduced Euclidean volume $\mathrm{vol}(\mathscr{O})$ | Euclidean lengths $\ell_\alpha$ and $\ell_\gamma$ of singular geodesics of $\mathscr{O}$ |
|---|---|---|
| $\alpha = \frac{2\pi}{3}$ | $\frac{32\sqrt{2}}{3} = 15.0849$ | $2\sqrt{6} = 4.89898$ |
| $X = \sqrt{\frac{2}{3}} = 0.8165$, $g = 1$ | $\frac{1}{45\sqrt{2}} = 0.01571$ | 20 |
| $\alpha = \frac{4\pi}{5}$ | 48.5817 | 9.61766 |
| $X = 0.811618$, $g = -0.6757$ | 0.008185 | 32.0835 |
| $\alpha = \frac{5\pi}{6}$ | 11.6288 | 11.814 |
| $X = 0.810809$, $g = -1$ | 0.006679 | 38.416 |
| $\alpha = \frac{19\pi}{20}$ | 834.486 | 41.1951 |
| $X = 0.809175$, $g = 0.527436$ | 0.00192325 | 127.838 |
| $\alpha = \pi - 0.02$ | 51707 | 324.897 |
| $X = 0.80902$, $g = 0.99163$ | 0.000243936 | 1004.06 |

—the reduced Euclidean volume $\mathrm{vol}(\mathscr{O}(\alpha,\alpha;\gamma)) = \frac{\mathrm{Vol}(\mathscr{O}(\alpha,\alpha;\gamma))}{d\ell_\alpha^2\ell_\gamma}$, where $d$ is the shortest distance between singular components; in our model $d = |Q_0Q_1| = 2$.

The table data is sorted in the decreasing order of the reduced Euclidean volume of the cone-manifold $\mathscr{O}(\alpha,\alpha;\gamma)$.

## *REFERENCES*


**1.** *P. Koebe* Über die Uniformisierung beliebiger analytischer Kurven // Nachr. Akad. Wiss. Gött., II. Math.-Phys. Kl. 1907. P. .

**2.** *Perelman G.* The entropy formula for the Ricci flow and its geometric applications // arXiv: math.DG/ 0211159[math.DG]. 2002.

**3.** *Perelman G.* Ricci flow with surgery on three-manifold // arXiv:math.DG/0303109[math.DG]. 2003.

**4.** *Perelman G.* Finite extinction time for the solutions to the Ricci flow on certain three-manifolds // arXiv:math.DG/0307245[math.DG]. 2003.

**5.** *Vesnin A. Yu. and Rasskazov A. A.* Isometries of hyperbolic Fibonacci manifolds // Sib. Math. J. 1999. V. 40, N 1. P. 9–22.

**6.** *Riley R.* An elliptical path from parabolic representations to hyperbolic structure // Topology of Low–Dimension Manifolds. : Springer-Verlag, 1979. P. 99–133. (Lect. Notes Math.; V. 722).

**7.** *Thurston W. P.* The Geometry and Topology of 3-Manifolds, Princeton Univ., Princeton (1980).

**8.** *Mednykh A. and Rasskazov A.* Volumes and degeneration of cone-structures on the figure-eight knot // Tokyo J. Math. 2006. V. 29, N 2. P. 445–464.

**9.** *Shmatkov R. N.* Properties of Euclidean Whitehead link cone-manifolds // Sib. Adv. Math. 2003. V. 13, N 1. P. 55–86.

**10.** *Sokolova D. Yu.* On the Euclidean structure on a trefoil knot with a bridge // Mat. Zametki YaGU. 2013. V. 20, N 1. P. 128–140.


*September 24, 2015*

A. D. Mednykh; D. Yu. Sokolova
Sobolev Institute of Mathematics
Novosibirsk State University, Novosibirsk, Russia
`mednykh@math.nsc.ru`

Yakutian Mathematical Journal
October—December, 2015. Vol. 22, No. 4

UDC 512.547.214

# FINITE GROUPS WITH IRREDUCIBLE CHARACTERS OF LARGE DEGREE

**S. S. Poiseeva**

**Abstract.** We study a nontrivial finite group $G$ with an irreducible complex character $\Theta$ satisfying $|G| \leq 2\Theta(1)^2$. We prove that in the case $\Theta(1) = p^2q$, where $p > q$ are distinct primes, $G$ is a solvable group with an abelian normal subgroup of index $p^2q$. Using the classification of finite simple groups, we prove that every finite simple nonabelian group with an abelian Sylow $p$-subgroup $P \neq 1$ of order at most $p^2$ such that $2|P|^3 > |G|$ is isomorphic to the group $L_2(q)$, where $q$ is either a prime or a prime square.

**Keywords:** finite group, character, irreducible character degree

## Introduction

Let $G$ be a finite group with an irreducible representation over the field of complex numbers with character $\Theta$.

In the general case irreducible character degrees hold rather scanty information on the structure of the group. Thus, it is natural to study the groups whose irreducible character degrees have additional properties and satisfy certain restrictions.

Refer to a finite group $G \neq 1$ possessing an irreducible complex character $\Theta$ with $2\Theta(1)^2 \geq |G|$ as an *$LC(\Theta)$-group*.

The goal of this article it to study finite $LC(\Theta)$-groups with $\Theta(1) = p^2q$, where $p$ and $q$ are distinct primes and $p > q$.

Since $|G| \leq 2p^4q^2 < 2p^6$, where $p > q$ are distinct primes, and the order of Sylow $p$-subgroups of $G$ is at least $p^2$, firstly we use the classification of finite simple groups to describe simple nonabelian groups $G$ with abelian Sylow $p$-subgroups $P$ of order $p^2$ such that $|G| < 2|P|^3$.

Vdovin proved [1] that $|A|^3 < |G|$ whenever $G$ is a finite simple group not isomorphic to $PSL_2(q)$ and $A$ is its abelian subgroup. Using the classification of finite simple groups, we obtain the following results.

**Theorem 1.** *Let $G$ be a finite simple nonabelian group with an abelian Sylow $p$-subgroup $P \neq 1$ of order at most $p^2$. If $2|P|^3 > |G|$ then $G$ is isomorphic to the group $L_2(q)$, where $q$ is either a prime or a prime square.*

**Theorem 2.** *Let $G$ be an $LC(\Theta)$-group. If $\Theta(1) = p^2q$, where $p > q$ are distinct primes, then $G$ is a $p$-solvable group.*

In the proof of Theorem 2 we also establish that for $|G| \notin \{486, 648\}$ the order $|G|$ equals $p^2q^bm$, where $(pq, m) = 1$. The following theorem describes $LC(\Theta)$-groups.

**Theorem 3.** *Let $G$ be an $LC(\Theta)$-group with $\Theta(1) = p^2q$, where $p$ and $q$ are distinct primes with $p > q$. Then $G$ has an abelian normal subgroup $M$ of index $p^2q$.*

Assume that all groups are finite. The letters $p$ and $q$ stand for distinct primes throughout. For the necessary background concerning ordinary and modular representations of finite groups, see [2, 3]. Denote the inner product of two characters $\chi$ and $\psi$ of a group $G$ by $\langle\chi, \psi\rangle$. In particular, if these characters are irreducible then $\langle\chi, \psi\rangle = \delta_{\chi,\psi}$ (where $\delta$ is the Kronecker symbol). Denote the set of irreducible characters of a group $G$ by $\mathrm{Irr}(G)$.

## 1. Auxiliary Results

Recall a fundamental result of Clifford on the restriction of an irreducible character $\chi$ of a group $G$ to its normal subgroup $N$.

**Lemma 1** (Clifford). *Let $N \lhd G$ and let $\chi \in \mathrm{Irr}(G)$. Let $\theta$ be an irreducible constituent of $\chi_N$ and suppose that $\theta = \theta_1, \theta_2, \dots, \theta_t$ are the distinct conjugates of $\theta$ in $G$. Then*

$$\chi_N = e\sum_{i=1}^{t}\theta_i,$$

*where $e = [\chi_N, \theta]$.*

Proof. See Theorem 6.2 of [3].

**Lemma 2.** *Let $\chi$ be an irreducible character of $G$. Let $N \lhd G$ and let $\theta$ be an irreducible constituent of $\chi_N$. If $I_G(\theta)/N$ is cyclic then $e_N(\chi) = 1$; i.e., $\chi_N = \sum_{i=1}^{t}\theta_i$.*

Proof. See 9.12 in [4].

Recall Ito's Theorem on the irreducible character degree (Theorem 6.15 of [3]).

**Lemma 3** (Ito). *Let $N \lhd G$ be abelian. Then $\chi(1)$ divides $|G : N|$ for all $\chi \in \mathrm{Irr}(G)$.*

Below we need the following two theorems of Zenkov [5].

**Lemma 4.** *Let $G$ be a finite group with a Sylow $p$-subgroup $P$. If $p \neq 2$ is not a Mersenne prime then $P \cap P^x = O_p(G)$ for some $x \in G$. If $O_p(G) = 1$ then $|P|^2 < |G|$.*

**Lemma 5.** *Let $G$ be a finite group with a Sylow $p$-subgroup $P$ and with solvable radical $S(G)$, and assume that $|G| = p^a m$ with $(p, m) = 1$, where $p$ is a prime. If $p^a \geq m$ then one of the following holds:*

*(1) $G$ includes a characteristic $p$-subgroup of order $> p^a m^{-1}$;*

*(2) one of the following holds in the quotient $\overline{S(G)} = S(G)/O_p(G)$:*

*(2a) $p = 2$, while $q = 2^n + 1$ is a Fermat prime, and $\overline{S(G)}$ includes a section isomorphic to $(Z_{2^n+1} \leftthreetimes Z_{2^n}) \wr Z_2$, for $n > 2$ and $((Z_3 \leftthreetimes Z_2) \wr Z_2) \wr Z_2$, $V_9 \leftthreetimes SD_{16}$, $(V_9 \leftthreetimes Z_8) \wr Z_2$, $(V_9 \leftthreetimes Q_8) \wr Z_2$ for $n = 1$;*

*(2b) $p = 2^n - 1$ is a Mersenne prime and $\overline{S(G)}$ includes a section isomorphic to $(Z_{2^n} \leftthreetimes Z_p) \wr Z_p$;*

*(2c) $p = 2$ and $S(G)$ includes a section isomorphic to $((V_{7^2} \leftthreetimes SD_{2^5}) \wr Z_2) \wr Z_2$; furthermore, in all three cases (2a), (2b), and (2c) the action is faithful;*

*(3) $p = 2$ and the quotient $\widetilde{G} = G/S(G)$ includes a section $L$ isomorphic to one of the groups $(((S_5 \wr Z_2) \wr Z_2) \wr Z_2) \wr Z_2$, $\mathrm{Aut}(A_6) \wr Z_2$, $\mathrm{Aut}(L_3(2)) \wr Z_2$, and*

$\mathrm{Aut}(L_3(4)) \wr Z_2$; *furthermore, each component of* $L$ *is an isomorphic image of the corresponding component of* $E(\widetilde{G})$.

Two auxiliary statements, which we use to study $LC(\Theta)$-groups with $\Theta(1) = p^2q$, are justified in [6].

**Lemma 6.** *If* $G$ *is an* $LC(\Theta)$*-group and* $M$ *is a proper normal subgroup then the character* $\Theta_M$ *is reducible.*

**Lemma 7.** *Let* $G$ *be an* $LC(\Theta)$*-group and let* $N$ *be the proper normal subgroup of* $G$*. If* $\Theta(1) = m$ *then* $(|G/N|, m) \neq 1$.

Let us present a result of [7] on the solvability of a group with a self-normalizing Sylow $p$-subgroup ($p > 3$).

**Lemma 8.** *If* $G$ *is a finite group with a Sylow* $p$*-subgroup* $P$ *such that* $p > 3$ *and* $N_G(P) = P$ *then* $G$ *is solvable.*

To estimate the orders of abelian subgroups of simple groups, we need the following result of Vdovin [1].

**Lemma 9.** *Let* $G$ *be a nonabelian finite simple group and* $G \not\cong L_2(q)$*, where* $q = p^t$ *for some prime* $p$*. If* $A$ *is an abelian subgroup of* $G$ *then* $|A|^3 < |G|$.

**Lemma 10.** *Let* $G = G'$ *be not a* $p$*-solvable group, with a Sylow* $p$*-subgroup* $P$ *of order* $p > 3$*, and not normal subgroups of order* $2$*. If* $|G| < p^3$ *then* $G \simeq L_2(r)$*, where* $r = p$ *or* $r = 2^a = p - 1$.

Proof. See Corollary 5.2, Chapter VIII of [8].

**Lemma 11** (Kazarin). *Let* $G$ *be a finite group and let* $x \in G$ *be an element of prime order* $q$*. If* $|G : C_G(x)|$ *is a power of a prime* $p$ *then* $\langle x^G\rangle' = O_p(\langle x^G\rangle)$*. In particular, the commutant of the normal closure of* $x$ *in* $G$ *is a* $p$*-group.*

Proof. See [9].

**Proposition 1.** *Let* $G$ *be a finite group and let* $x$ *be an element of* $G$ *such that* $|G : C_G(x)|$ *is a power of a prime* $p$*. Then* $[x^G, x^G] \subseteq O_p(G)$.

Proof. See [10].

## 2. Proof of Theorem 1

By the classification of finite simple nonabelian groups (see [11] for more details), all simple nonabelian groups belong to the following families:

I. The classical simple groups of Lie type: $L_n(q)$ for $n \geq 2$; $U_n(q)$ for $n \geq 3$; $S_{2n}(q)$ for $n \geq 2$; $P\Omega_{2n+1}(q)$ for $n \geq 2$; $P\Omega^{\pm}_{2n}(q)$ for $n \geq 4$.

II. The exceptional simple groups of Lie type: $G_2(q)$, $F_4(q)$, $E_6(q)$, $E_7(q)$, $E_8(q)$, ${}^2G_3(q)$, $({}^2F_4(q))'$, ${}^3D_4(q)$, and ${}^2B_2(q)$, where $q$ is a prime power.

III. The sporadic simple groups.

IV. The alternating groups $A_n$ for $n \geq 5$.

We prove Theorem 1 in steps for all listed groups. Recall that in Theorem 1 the Sylow $p$-subgroup $P \neq 1$ is of order at most $p^2$.

I. The classical simple groups of Lie type.

While studying Chevalley groups, we denote the order $q$ field by $GF(q)$, and its characteristic by $r$.

(a) Suppose that $G \cong L_n(q)$ with $n \geq 2$. It is easy to see that $|G| > q^{n^2-n}$, while the greatest prime divisor of the group order is at most $q^n - 1 < q^n$. Since

$2p^6 < p^{7n}$, it follows that $q^{7n} > q^{n^2-n}$. Hence, $n^2 - n < 7n$ and $n^2 - 8n < 0$, and so $n < 8$.

For $n = 7$ we have

$$|G| = 1/dq^{21}(q^7 - 1)(q^6 - 1)(q^5 - 1)(q^4 - 1)(q^3 - 1)(q^2 - 1) > q^{42},$$

where $d = (7, q-1)$. Since $q^7 - 1 = (q-1)(q^6 + q^5 + q^4 + q^3 + q^2 + q + 1)$, the greatest prime divisor of the group order is at most $q^6 + q^5 + q^4 + q^3 + q^2 + q + 1 < 2q^6$. If $p^2 | q^6 + q^5 + q^4 + q^3 + q^2 + q + 1$ then $p^2 < q^6 + q^5 + q^4 + q^3 + q^2 + q + 1 < 2q^6$.

Hence, $2(2q^6)^3 = 2^4q^{18} > |G| > q^{42}$, which is impossible.

If $p^2 \nmid q^6 + q^5 + q^4 + q^3 + q^2 + q + 1$ then $p \mid q^4 + q^3 + q^2 + q + 1$ because $q^5 - 1 = (q-1)(q^4 + q^3 + q^2 + q + 1)$. Since $q^4 + q^3 + q^2 + q + 1 < 2q^4$ and $q^5 - 1$ are coprime to all divisors of the group order of the form $q^i - 1$ for $i \neq 5$, it follows that $p^2 < 2q^4$ and $p^2 \mid q^5 - 1$. Hence, $2p^6 < 2(q^5 - 1)^3 < 2q^{15} < q^{42}$, which excludes this case.

Suppose that $n = 6$, and so $|G| > q^{30}$. Since $q^5 - 1 = (q-1)(q^4 + q^3 + q^2 + q + 1)$, the greatest prime divisor of the group order is at most $q^4 + q^3 + q^2 + q + 1 < 2q^4$. If $p^2 \mid q^4 + q^3 + q^2 + q + 1$ then $p^2 < q^4 + q^3 + q^2 + q + 1 < 2q^4$. Hence, $2(2q^4)^3 = 2^4q^{12} > |G| > q^{30}$, which is false.

If $p^2 \nmid q^4 + q^3 + q^2 + q + 1$ then $p \mid q^2 + q + 1$ because $q^6 - 1 = (q^3 - 1)(q^3 + 1) = (q-1)(q^2 + q + 1)(q + 1)(q^2 - q + 1)$. Then $p < q^2 + q + 1$, and since $q^2 + q + 1 < 2q^2$ in the factorization of the group order is of degree 2, we have $p^2 < (2q^2)^2$. Hence, $2p^6 < 2(2q^2)^6 = 2^7q^{12} < q^{30}$, which excludes this case.

For $n = 5$ the greatest prime divisor of the group order is also at most $q^4 + q^3 + q^2 + q + 1 < 2q^4$, and $|G| > q^{20}$. If $p^2 | q^4 + q^3 + q^2 + q + 1$ then $p^2 < q^4 + q^3 + q^2 + q + 1 < 2q^4$. Hence, $2(2q^4)^3 = 2^4q^{12} > |G| > q^{20}$, which is false.

If $p^2 \nmid q^4 + q^3 + q^2 + q + 1$ then $p \mid q^2 + q + 1$ because $q^3 - 1 = (q-1)(q^2 + q + 1)$. Since $q^2 + q + 1 < 2q^2$ is coprime to all divisors of the group order of the form $q^i - 1$ for $i \neq 3$, it follows that $p^2 < 2q^2$ and $p^2 | q^3 - 1$. Hence, $2p^6 < 2(q^3 - 1)^3 < q^{20}$, which excludes this case as well.

If $n = 4$ then the greatest prime divisor of the group order is at most $q^2 + q + 1 < 2q^2$, and $|G| > q^{12}$. If $p^2 \mid q^2 + q + 1$ then $p^2 < q^2 + q + 1 < 2q^2$. Hence, $2(2q^2)^3 > |G| > q^{12}$, which is false. If $p^2$ does not divide $q^2 + q + 1$ then $p \mid q^2 + 1$. Since $q^2 + 1$ is coprime to all divisors of the group order of the form $q^i - 1$ for $i \neq 3$, it follows that $p^2 | q^2 + 1$. So, $2p^6 < 2(q^2 + 1)^3 < q^{12}$, which excludes this case as well.

If $n = 3$ then $|G| = (1/d)q^3(q^3 - 1)(q^2 - 1)$, where $d = (3, q-1)$. Therefore, $p^2$ divides either $(q-1)^2$ or $q + 1$ or $q^2 + q + 1$. Straightforward calculations exclude this case.

Thus, if $G \simeq L_n(q)$ satisfies the hypotheses of Theorem 1 then $n = 2$.

(b) Suppose that $G \simeq U_n(q)$ with $n \geq 3$. It is easy to see that $|G| > q^{n^2-n}$, while the greatest prime divisor of the group order is at most $q^n$. Since $2p^6 < p^{7n}$, it follows that $q^{7n} > q^{n^2-n}$. Hence, $n^2 - n < 7n$ and $n^2 - 8n < 0$. Therefore, $n < 8$.

Suppose that $n = 7$. Then

$$|G| = 1/dq^{21}(q^7 + 1)(q^6 - 1)(q^5 + 1)(q^4 - 1)(q^3 + 1)(q^2 - 1) > q^{42},$$

where $d = (7, q+1)$. Since $q^7 + 1 = (q+1)(q^6 - q^5 + q^4 - q^3 + q^2 - q + 1)$, the greatest prime divisor of the group order is at most $q^6 - q^5 + q^4 - q^3 + q^2 - q + 1 < q^6$. If $p^2 \mid q^6 - q^5 + q^4 - q^3 + q^2 - q + 1$ then $p^2 < q^6 - q^5 + q^4 - q^3 + q^2 - q + 1 < q^6$. Hence, $2(q^6)^3 > |G| > q^{42}$, which is false. If $p^2 \nmid q^6 - q^5 + q^4 - q^3 + q^2 - q + 1$ then $p | q^4 - q^3 + q^2 - q + 1$ because $q^5 + 1 = (q+1)(q^4 - q^3 + q^2 - q + 1)$. Since

$q^4 - q^3 + q^2 - q + 1 < q^4$ is coprime to all divisors of the group order of the form $q^i - (-1)^i$, it follows that $p^2 < q^4$ and $p^2 | q^5 + 1$. Hence, $2p^6 < 2(q^5+1)^3 < q^{42}$, which excludes this case.

Suppose that $n = 6$, and so $|G| > q^{30}$. Since $q^5+1 = (q+1)(q^4-q^3+q^2-q+1)$, the greatest prime divisor of the group order is at most $q^4 - q^3 + q^2 - q + 1 < q^4$. If $p^2 | q^4 - q^3 + q^2 - q + 1$ then $p^2 < q^4 - q^3 + q^2 - q + 1 < q^4$. Hence, $2(q^4)^3 = 2q^{12} > |G| > q^{30}$, which is false. If $p^2 \nmid q^4 - q^3 + q^2 - q + 1$ then $p | q^2 + q + 1$ because $q^3 - 1 = (q-1)(q^2+q+1)$.

Therefore, $p < q^2 + q + 1$. Since $q^2 + q + 1$ is coprime to all divisors of the group order of the form $q^i - (-1)^i$ for $i \neq 3$, it follows that $p^2 | (q^2 + q + 1)$. Consequently, $2p^6 < 2(q^2+q+1)^3 < q^{30}$, which excludes this case as well.

For $n = 5$ the greatest prime divisor of the group order is also at most $q^4 - q^3 + q^2 - q + 1 < q^4$. If $p^2 | q^4 - q^3 + q^2 - q + 1$ then $p^2 < q^4 - q^3 + q^2 - q + 1 < q^4$. Hence, $2(q^4)^3 > |G| > q^{20}$, which is impossible. If $p^2$ does not divide $q^4 - q^3 + q^2 - q + 1$ then $p | q^2 + 1$. Since $q^2 + 1$ is coprime to all divisors of the group order, we have $p^2 | q^2 + 1$. Then $2p^6 < 2(q^2+1)^3 < q^{20}$, which is excluded.

If $n = 4$ then $|G| > q^{12}$ and the greatest prime divisor of the group order is at most $q^2 + 1$. If $p^2 \mid (q^2+1)$ then $p^2 < q^2 + 1$. Hence, $2(q^2+1)^3 > |G| > q^{12}$, which is false. If $p^2$ does not divide $(q^2+1)$ then $p | q^2 - q + 1$ because $q^3 + 1 = (q+1)(q^2-q+1)$. Since $q^2 - q + 1 < q^2$ is coprime to all divisors of the group order of the form $q^i - (-1)^i$, it follows that $p^2 < q^2$ and $p = q$. Hence, $2p^6 < q^{12}$, which excludes this case as well.

If $n = 3$ then $|G| = (1/d)q^3(q^3+1)(q^2-1)$, where $d = (3, q+1)$. Therefore, $p^2$ divides either $q - 1$ or $(q+1)^2$ or $q^2 - q + 1$. Straightforward calculations exclude this case.

(c) Suppose that $G$ is isomorphic to $S_{2n}(q)$ or $P\Omega_{2n+1}(q)$ with $n \geq 2$. The order of $G$ equals

$$1/dq^{n^2}(q^{2n}-1)\dots(q^2-1),$$

where $d = (2, q-1)$. If $n \geq 2$ then $|G| > q^{2n^2}$.

The greatest prime divisor of the order of $G$ is at most $q^n + 1$. If $p^2 | (q^n + 1)$ then $2(q^n+1)^3 < q^{3n+2} < q^{2n^2}$ with $n \geq 3$.

For $(q, n) = (2, 2)$ the group $G \simeq S_4(2)$ is not simple and $S_4(2)' \simeq L_2(9)$.

If $p^2 \nmid (q^n + 1)$ then $p | q^{n-1} + 1$; hence, $p^2 | q^{n-1} + 1$. Therefore, $2p^6 < 2(q^{n-1} + 1)^3 < q^{3n-1} < q^{2n^2}$ for $n \geq 3$, which excludes this case.

(d) Suppose that $G \simeq P\Omega^{\pm}_{2n}(q), n \geq 4$.

It is easy to see that $|G| > q^{2n^2-2n}$, while the greatest prime divisor of the order is at most $q^n + 1$. If $p^2 | (q^n + 1)$ then $2(q^n+1)^3 < q^{3n+2}$ for $n \geq 4$, and so $2p^6 < |G|$.

If $p^2 \nmid (q^n + 1)$ then $p | q^{n-1} + 1$; hence, $p^2 | q^{n-1} + 1$. Therefore, $2p^6 < 2(q^{n-1} + 1)^3 < q^{3n-1} < q^{2n^2-2n}$ with $n \geq 4$, which excludes this case.

II. The exceptional simple groups of Lie type.

(a) Suppose that $G \simeq G_2(q)$. The order of $G$ equals $q^6(q^6-1)(q^2-1)$, and so the greatest prime divisor of the group order is at most $q^2 + q + 1 < 2q^2$. If $p^2 | q^2 + q + 1$ then $p^2 < q^2 + q + 1 < 2q^2$. Hence, $2(2q^2)^3 = 2^4q^6 > |G|$. But $|G| > q^{13}$, so that this case is excluded.

If $p^2 \nmid q^2 + q + 1$ then $p | q^2 - q + 1$. Since $q^2 - q + 1$ is coprime to all divisors of the group order of the form $q^i \pm 1$ for $i \neq 3$, it follows that $p^2 | q^2 - q + 1$. Then $p^2 < q^2 - q + 1 < q^2$, and so $2p^6 < 2q^6 < q^{13}$, which excludes this case.

(b) Suppose that $G \simeq F_4(q)$, which is of order $q^{24}(q^{12}-1)(q^8-1)(q^6-1)(q^2-1)$. The greatest prime divisor of the order of $G$ is at most $q^4+1$. Since $2(q^4+1)^6 < |G|$, this case is excluded.

(c) Suppose that $G$ is one of the groups $E_6(q)$, $E_7(q)$, $E_8(q)$, or ${}^2E_6(q)$. The orders of all these groups are greater than $q^{72}$, while the greatest prime divisor of the order is at most $q^9$. Consequently, these groups are also excluded.

(d) Suppose that $G \simeq^2 G_2(q)$, $q^3(q^3+1)(q-1)$. Furthermore, $q = 3^{2n+1}$ with $n \geq 1$, while the greatest prime divisor $p$ of the group order is at most $q+\sqrt{3q}+1 < 2q$. Suppose that $q > 3$. If $p^2|q+\sqrt{3q}+1$ then $p^2 < q+\sqrt{3q}+1 < 2q$. Hence, $2(2q)^3 = 2^4q^3 < |G|$, which is impossible. If $p^2 \nmid q+\sqrt{3q}+1$ then $p = q$ is the greatest prime divisor of the order of $G$. Then $2q^6 \geq |G|$, so that $q = 3$ and $G \simeq L_2(8).3$, which is not a simple group.

(e) Suppose that $G \simeq ({}^2F_4(q))'$, which is of order $q^{12}(q^6+1)(q^4-1)(q^3+1)(q-1) > 2q^{16}$, where $q = 2^{2n+1}$ and $n \geq 2$, while the greatest prime divisor $p$ of the group order is at most $q^4-q^2+1 < q^4$. If $p^2|q^4-q^2+1$ then $p^2|q^4-q^2+1 < q^4$. Hence, $2(q^4)^3 = 2q^{12} < 2q^{16} < |G|$. If $p^2 \nmid q^4-q^2+1$ then $p|q^2+1$. Hence, $p < (q^2+1)$, and since $q^2+1$ in the factorization of the group order is of degree 2, it follows that $p^2 < (q^2+1)^2$. Therefore, $2p^6 < 2(q^2+1)^6 < |G|$, which excludes this case as well.

(f) Suppose that $G \simeq^3 D_4(q)$ which is of order $q^{12}(q^8+q^4+1)(q^6-1)(q^2-1) > q^{26}$, while the greatest prime divisor $p$ is at most $q^4-q^2+1 < q^4$. If $p^2|q^4-q^2+1$ then $p^2 < q^4-q^2+1 < q^4$. Hence, $2(q^4)^3 = 2q^{12} < |G|$.

If $p^2 \nmid q^4-q^2+1$ then $p|q^2+q+1$. Hence, $p < (q^2+q+1)$, and since $q^2+q+1 < 2q^2$ in the factorization of the group order is of degree 2, it follows that $p^2|q^2+q+1$. Therefore, $p^2 < q^2+q+1 < 2q^2$, and so $2p^6 < 2(2q^2)^3 = 2^4q^6 < |G|$, which excludes this case as well.

(g) Suppose that $G \simeq^2 B_2(q) \simeq Sz(q)$ which is of order $q^2(q^2+1)(q-1)$, where $q = 2^{2n+1}$ and $n \geq 1$. The greatest prime divisor $p$ of the order of $G$ is at most $q+\sqrt{2q}+1 < 2q$. If $p^2|q+\sqrt{2q}+1$ then $p^2 < q+\sqrt{2q}+1 < 2q$. Hence, $2(2q)^3 < |G|$. If $p^2 \nmid q+\sqrt{2q}+1$ then $p|q-\sqrt{2q}+1$, whence $p^2|q-\sqrt{2q}+1$ and $p^2 < q-\sqrt{2q}+1 < q$. Therefore, $2p^6 < 2q^3 < |G|$, which excludes this case as well.

III. The sporadic simple groups. All cases are excluded (see [12]).

IV. The alternating groups $A_n$ with $n \geq 5$.

Since $|A_n| = n!/2 \leq 2n^6$, it follows that $n! \leq 4n^6$. On the other hand, it is known that $n! \geq n^{n/2}$. Therefore, $n^{n/2} \leq 4n^6 < n^7$. Hence, $n/2 < 7$ and $n < 14$. For $n = 5, 6$ the groups $A_5$ and $A_6$ of orders 60 and 360 appear, which have abelian Sylow $p$-subgroups of orders 4 and 9 respectively. Observe that $A_5 \simeq L_2(5)$ and $A_6 \simeq L_2(9)$. All remaining possibilities are excluded.

The proof of the theorem is complete.

## 3. Proof of Theorem 2

Suppose that $G$ is an $LC(\Theta)$-group with $\Theta(1) = p^2q$, where $p$ and $q$ are distinct primes with $p > q$. Fix this notation.

Since $\sum_{\chi\in\mathrm{Irr}(G)} \chi(1)^2 = |G|$ and $2\chi(1)^2 \geq |G|$, the character $\Theta$ is a unique irreducible character of $G$ of the greatest degree; moreover, all values $\Theta(g)$ for $g \in G$ are integer rational numbers.

As [13] established, each irreducible character of an $LC(\Theta)$-group whose order is not a power of 2 is a constituent of the character $\Phi = \Theta^2$. Observe that $Z(G) = 1$ and $\Theta$ is faithful.

Take $P \in \mathrm{Syl}_p(G)$. Since $G$ is an $LC(\Theta)$-group with $\Theta(1) = p^2q$, where $p$ and $q$ are distinct primes with $p > q$, and furthermore, $p^4q^2 < |G| \leq 2p^4q^2$ and $pq||G|$, it follows that $|G| = p^aq^bm$ with $a, b, m \in \mathbb{N}$ and $2 \leq a$; i.e. the Sylow $p$-subgroups of $G$ are of order at least $p^2$. Observe that $|G| \leq 2p^4q^2 < 2p^6$ because $p > q$.

Let us state some upper bounds on the order of Sylow $p$-subgroups of $LC(\Theta)$-groups with $\Theta(1) = p^2q$ as the next lemma.

**Lemma 12.** *Suppose that $G$ is an $LC(\Theta)$-group with $\Theta(1) = p^2q$, where $p$ and $q$ are distinct primes with $p > q$. Then the Sylow $p$-subgroup $P$ of $G$ is of order at most $p^5$. If $P \triangleleft G$ then $|G| = 486$. If $P$ is not normal in $G$ then $|P| \leq p^3$.*

PROOF. Suppose that $G$ is an $LC(\Theta)$-group with $\Theta(1) = p^2q$, where $p > q$. Since $|G| \leq 2p^4q^2 < p^7$, it follows that $|P| \leq p^6$; otherwise, $q$ does not divide $|G|$. If $|P| = p^6$ then $2p^4q^2 < 2p^6$ because $q < p$. Hence, $|G| = |P| = p^6$, which is false, as in this case $q$ does not divide $|G|$. Hence, $|P| \leq p^5$.

Suppose that $|P| = p^5$. Since $|G| = p^5q^bm \leq 2p^4q^2$, where $(m, pq) = 1$, it follows that $b \leq 2$. But $p > q \geq 2$. Consequently, we may assume that $b = 1$. In this case $pm \leq 2q$. Since for $m = 1$ we have $P \triangleleft G$, it follows that $m \geq 2$ and $p < q$ contrary to the assumption. Therefore, $|P| \triangleleft G$ and $|G| = p^5q$.

Furthermore, $|Z(P)| = p$ and the order $q$ subgroup $Q$ of $G$ acts nontrivially on $Z(P)$. Hence, $q \mid p - 1$. There exist $p - 1$ conjugate characters of $G$ of degree $p^2q$ vanishing on $P \setminus Z(P)$. Therefore, $Q$ partitions them into length $q$ orbits. Thus, we have $(p - 1)/q$ characters of $G$ of degree $p^2q$. Since only one degree $p^2q$ character is possible, it follows that $q = p - 1$. Hence, $p = 3$, $q = 2$, and $|G| = 486$.

Suppose that $P \triangleleft G$ and $|P| = p^4$. By Lemma 1,

$$\Theta_P = e\sum_{i=1}^{s} \chi_i,$$

where $e, s \mid |G/P|$ and $\chi_i \in \mathrm{Irr}(P)$. Since the order of $G/P$ is coprime to $p$ and $\Theta(1) = p^2q$, it follows that $\chi_i(1) = p^2$, which leads to a contradiction because the sum of the squared degrees of the irreducible characters of $P$ in this case is greater than $p^4$. Therefore, in the case that $P$ is normal its order equals $p^5$.

The proof of Lemma 12 is complete.

The next lemma shows that, when $\Theta(1) = 3^2 \cdot 2$, the order of $G$ lies in $\{342, 486, 504, 648\}$.

**Lemma 13.** *Suppose that $G$ is an $LC(\Theta)$-group with $\Theta(1) = p^2q$, where $p$ and $q$ are distinct primes with $p > q$. Then either $p > 3$ or $|G| \in \{342, 486, 504, 648\}$.*

PROOF. Suppose that $p = 3$. Then $q = 2$. Therefore, $|G| \leq 3^4 \cdot 2^3 = 648$, and furthermore, the order of $G$ is divisible by $\Theta(1) = 3^2 \cdot 2 = 18$.

Considering that

$$\frac{|G|}{|\Theta(1)|} = \frac{2^3 \cdot 3^4}{2 \cdot 3^2} \leq 2^2 \cdot 3^2 = 36,$$

and then that a nonsolvable group of order at most $81 \cdot 8 = 648$ has a nonsolvable composition factor contained in [12], we exclude the factors of this type. Hence, we conclude that $G$ is solvable and $|G|$ divides one of the numbers $648 = 2^3 \cdot 3^4$, $342 = 19 \cdot 2 \cdot 3^2$, $486 = 2 \cdot 3^5$, and $504 = 2^3 \cdot 3^2 \cdot 7$. Applying the GAP [14], we verify that all these possibilities are realized.

Assume henceforth that $p > 3$. Verify that the order of $P$ can be equal to $p^3$ only if $O_p(G) > 1$.

**Lemma 14.** *If $O_p(G) = 1$ then either $P = N_G(P)$ and $|P| = p^3$, or $|G| = p^2q^bm$, where $(pq, m) = 1$.*

PROOF. Suppose that $O_p(G) = 1$. Since $|G|$ is divisible by $\Theta(1) = p^2q$, it follows that $|G| = p^aq^bm$, where $(pq, m) = 1$ and $a, b, m \in \mathbb{N}$. Lemmas 12 and 13 yield $p > 3$ and $2 \le a \le 3$. Lemma 4 shows that $p^a < q^bm$, and from $2\Theta(1)^2 \ge |G|$ we infer that $q^{b-2}m \le 2p^{4-a}$.

Suppose that $a = 3$ and $|G| = p^3q^bm$, where $p^3q^bm \le 2p^4q^2$. Then $q^{b-2}m \le 2p$. Since $O_p(G) = 1$, Lemma 4 yields $p^3 < q^bm \le 2pq^2$. Hence, $p^2 \le 2q^2$.

Suppose that $|N_G(P)| = n|P| = np^3$, where $n > 1$. Since $P \cap P^x = 1$ for at least one $x \in G$ by Lemma 4, it follows that

$$|G| \ge |N_G(P)| + \frac{n^2|P|^2}{d} = np^3 + \frac{n^2p^6}{d},$$

where $d = |N_G(P) \cap N_G(P)^x| | n$. Therefore, $p^3q^bm \ge np^3 + np^6$, whence $q^bm > np^3$. Since $np^3 < q^bm \le 2pq^2$, it follows that $np^2 < 2q^2$, where $n > 1$; i.e. $p^2 < q^2$, which is impossible.

Thus, for $O_p(G) = 1$ we have $N_G(P) = P$ for $|P| = p^3$ or $|G| = p^2q^bm$, where $(pq, m) = 1$.

**Lemma 15.** *If $G$ is not $p$-solvable then $O_p(G) = 1$.*

PROOF. Suppose that $O_p(G) > 1$. By Lemma 3, all irreducible character degrees of $G$ divide $|G : O_p(G)|$ whenever $O_p(G)$ is abelian. Since $G$ is not $p$-solvable, it follows that $P$ is not normal in $G$. Therefore, either $|O_p(G)| = p$ or $|O_p(G)| = p^3$. Verify that in all cases $G$ has a normal subgroup $U$ of order $p$. If $|O_p(G)| = p$ then this holds.

By Lemma 3, if $|O_p(G)| = p^3$ then $P$ is not abelian, and so $|Z(P)| = p$ and $U = Z(P) \lhd G$. Put $H = C_G(U)$. Since $U \lhd G$, it follows that $H \lhd G$ and $G/H \le \mathrm{Aut}(U)$, and so it is isomorphic to a cyclic group of order dividing $p - 1$. Since $Z(G) = 1$, it follows that $G/H \ne 1$ and it is a cyclic group. Lemma 7 implies that $p - 1$ is a power of $q$. Clifford's Theorem (Lemma 1) yields

$$\Theta_H = e\sum_{i=1}^{s} \chi_i,$$

where $\chi_i$ are conjugate irreducible characters, while $e$ and $s$ divide $|G/H| = q^\lambda$. Hence, $p^2q = \Theta(1) = es\chi_i(1)$. Lemma 2 implies that $e = 1$ and $\chi_i(1) = p^2$, $s = q$. Thus, $H$ has $q$ irreducible characters $\chi_i$ of degree $p^2$; whence, $|H| > p^4q$ and $|G| > p^4qq^\lambda$. Therefore, $\lambda = 1$. Since $|G| = |H|q < 2p^4q^2$, it follows that $|P|q^{b-1}m = |H| < 2p^4q$.

If $|P| = p^4$ then $q^{b-1}m < 2q$. Furthermore, $|H : P| = mq^{b-1} < 2q$. For $b \ge 2$ we infer that $m \le 2$ and $H$ is solvable, and so is $G$. If $b = 1$ then $|H : P| = m < 2q$. Recall that $q \mid p - 1$. Since $p$ is odd, it follows that $q \le (p-1)/2$. Therefore, $|H : P| < p$ and $P \lhd H \lhd G$. Since $P$ is a characteristic subgroup, $P \lhd G$ and $G$ is $p$-solvable contrary to the assumption. Thus, the case $|P| = p^4$ is impossible.

Now $|P| = p^3$ and $|H : P| = q^{b-1}m \le 2pq$. For $b = 2$ we have $|H : P| = qm \le 2qp$, and so $m \le p$. Then $P \lhd H$, whence $P \lhd G$ and $G$ is $p$-solvable contrary to the assumption. Finally, for $b = 1$ we have $|H| = p^3m$, where $m \le 2pq$.

Observe that $|O_p(G)| = p$ and $m \le 2pq \le p(p-1)$. Thus, in the group $\bar{H} = H/O_p(G)$, whose order is less than $|P|^2$, the subgroup $O_p(\bar{H})$ is nontrivial by Lemma 4; this is a contradiction.

The proof of Lemma 15 is complete.

**Lemma 16.** *If $P = N_G(P)$ then $G$ is a $p$-nilpotent group of order $p^2q^bm$.*

Proof. Lemma 8 shows that $G$ is a solvable group. If $|G| = p^2q^bm$, where $(pq, m) = 1$, then $G$ is $p$-nilpotent by Burnside's Theorem.

Lemmas 12 and 13 imply that $|G| = p^aq^bm$, where $(pq, m) = 1$ and $2 \le a \le 3$. Suppose that $a = 3$, and furthermore, $q^{b-2}m < 2p$.

Suppose that $1 \ne m \le p$. If $m \mid |O_{p'}(G)|$ then for every prime divisor $r$ of $m$ the group $O_{p'}(G)P$ has a Hall $\{r, p\}$-subgroup $RP$, which is nilpotent, which contradicts the assumption. Similarly, there exists no normal subgroup $K$ of $G$ whose order is divisible by $r$ but not by $|P|$. Since $G$ is solvable, it is also $r$-solvable. Hence, $O_{r',r}(G)$ includes $P$ and so $N_G(P) \ne P$ by Frattini's argument; this is a contradiction. Therefore, either $m = 1$ or $m > p$ and it is prime.

If $m > p$ is a prime then $q^{b-2} < 2$, whence $1 \le b \le 2$. For $b = 1$ we arrive at a contradiction by analogy with the previous case. Therefore, $b = 2$. Since $P = N_G(P)$, it follows that for $m \mid |O_{p'}(G)|$ the subgroup $M$ of order $m$ is normal in $O_{p'}(G)$, and so in $G$ as well. In this case $H = C_G(M) \lhd G$ and $|G/H| \mid |\operatorname{Aut}(M)| = m - 1$. Since $m > p$ and $m < 2p$, it follows that $p$ does not divide $|G/H|$. At the same time, $H \ne G$ because $Z(G) = 1$; this is a contradiction. Consequently, $m$ does not divide $|O_{p'}(G)|$.

Thus, either $O_{p'}(G) = 1$ or $q^2 = |O_{p'}(G)|$.

Suppose that $|O_{p'}(G)| = q^2$. If $Q \in \operatorname{Syl}_q(G)$ is cyclic then the order of $G/C_G(Q)$ divides $q - 1$ and is not equal to 1. As in the previous case, this leads to a contradiction. Thus, $O_{p'}(G) = 1$.

Theorem 6.3.3 of [2] yields $C_G(O_p(G)) \subseteq O_p(G)$. If $|O_p(G)| = p$ then $|G| \mid p(p-1)$, which is a contradiction. If $|O_p(G)| = p^3$ then $P \lhd G$ contrary to the assumption. But if $|O_p(G)| = p^2$ then we arrive at a contradiction with Lemma 3.

Thus, we may assume that $m = 1$ and $|G| = p^3q^b$, where $q^{b-2} \le 2p$. From $|P| = |N_G(P)| = p^3$ we infer that the case $|O_p(G)| \ne 1$ leads to a contradiction. Theorem 6.3.3 of [2] and Lemma 4 imply that $|O_q(G)| = q^b$ and $p^2 < 2q^2$.

In particular, $G$ is a $p$-nilpotent group. Observe that $Q = O_q(G)$ is an elementary abelian group by Lemma 4 (otherwise some order $p$ element induces a trivial automorphism on $Q/\Phi(Q)$). Lemma 3 implies that $G$ lacks irreducible characters of degree $p^2q$; this is a contradiction.

All cases are considered and the proof of Lemma 16 is complete.

Assume henceforth that $|G| = p^2q^bm$, where $m$ is coprime to $pq$ and $b \in \mathbb{N}$.

**Lemma 17.** *Suppose that $G$ is an $LC(\Theta)$-group which is not $p$-solvable. Then $G$ is not simple and every simple section $L$ of $G$ which is not $p$-solvable is isomorphic to $L_2(r)$, where $r$ is a prime power; moreover, $r = p$, $p^2$, or $p - 1$.*

Proof. Take a simple $LC(\Theta)$-group $G$. Lemma 15 yields $O_p(G) = 1$, and by Lemmas 14 and 16 the order $|G| = p^2q^bm$, where $p > q$, is $(m, pq) = 1$.

Recall that $\Theta(1) = p^2q$. By Theorem 1, the group $G$ is isomorphic to $L_2(r)$. Since the order of the Sylow $p$-subgroups of $G$ divides $p^2$, it follows that $r$ equals $p$, $p^2$, or $p - 1$.

The character tables of these groups are available [15]. None of the groups listed is an $LC(\Theta)$-group for any choice of irreducible character $\Theta$.

Suppose that $G$ has a section $S$ isomorphic to a simple nonabelian group. Since the order of $S$ equals $p^aq^cm'$, where $a \le 2$, $m' \le m$, and $c \le b$, while $|G| < 2|P|^3$, Theorem 1 implies that $S \simeq L_2(r)$, where $r$ is a prime power.

The proof of Lemma 17 is complete.

We now finish the proof of Theorem 2.

Proof of Theorem 2. Lemma 4 and the conditions on $G$ yield

$$p^4q^2 < |G| = p^a q^b m \le 2p^4q^2,$$

where $a \le 4$ by Lemma 12. Since $G$ is not $p$-solvable, Lemmas 15 and 16 imply that $a = 2$, whence $q^{b-2}m \le 2p^2$ and $q^{b-2}m > p^2$.

Verify that the composition series of $G$ cannot contain two composition factors which are not $p$-solvable. Indeed, by Lemma 17 each factor is isomorphic to the group $L_2(r)$, where $r = p$, $p^2$, or $p-1$. Suppose that $G$ has two composition factors isomorphic to $L_2(p)$. Then

$$|G| \ge (1/4)p^2(p^2-1)^2,$$

while $|G| < 2p^4q^2$. Since $p > 3$, it follows that $q \le p-1$. Hence,

$$\frac{8p^4q^2}{p^2(p^2-1)^2} = \frac{8p^2q^2}{(p^2-1)^2} \le \frac{2p^2}{(p+1)^2} < 8.$$

Considering that $Z(G) = 1$, while the outer automorphism group of the group $L_2(p)$ is of order 2 for odd $p$, we conclude that $G$ has an index $\le 4$ subgroup isomorphic to $L_2(p) \times L_2(p)$.

Since for $r = p-1$ and $p > 3$ the number $r$ is a power of 2, we have $|L_2(p-1)| = p(p-1)(p-2) > |L_2(p)|$. Therefore, the case that at least one of the groups $L_2(p)$ is replaced by $L_2(p-1)$ is also excluded.

Let us show that the existence of a composition factor of $G$ isomorphic to $L_2(p^2)$ is also impossible. Indeed,

$$|L_2(p^2)| = (1/2)p^2(p^4-1) > |L_2(p) \times L_2(p)|.$$

Lemma 15 yields $O_p(G) = 1$. Put $M = O_{p'}(G)$. If $C_G(M) \subseteq M$ then Lemma 5 implies that $|M| \ge |P| + 1 = p^2 + 1$. Since $|P| = p^2$, we obtain

$$2p^4q^2 \ge |G| \ge |M||L_2(p)|p = (1/2)(p^2+1)p^2(p^2-1).$$

Observe that $q$ divides $|L_2(p)|$. Therefore, $q \le (p+1)/2$ and so $G$ coincides with a subgroup of order $(1/2)(p^2+1)p^2(p^2-1)$. Furthermore, in this case $|M| = p^2+1$. The latter is possible only in the case that $P$ has a unique orbit on $M \setminus \{1\}$. Then $MP$ is a Frobenius group and $P$ is a cyclic group, which is false.

Thus, $C_G(M) \lhd G$ is not $p$-solvable, but the order of $G$ is divisible by $|M|p|L_2(p)|$.

Moreover, there is a unique quotient of the group $C_G(M)$ which is not $p$-solvable, and it is isomorphic to $L_2(p)$. Therefore, $T \simeq L_2(p)$ or $T \simeq SL_2(p)$ is normal in $G$. There is a quotient of $G$ isomorphic to $C_p$, the cyclic group of order $p$, and the greatest normal $p$-solvable subgroup $S$ of $G$ is of $p$-length 1. It is easy to see that $G/S$ is isomorphic to $L_2(p)$, the automorphism group of $L_2(q)$ of order $p(p^2-1)$. Thus, we have the following possibilities:

(a) $G/S \simeq PGL_2(p)$, $q = 2$, and $|G| \le 8p^4$;

(b) $G/S \simeq L_2(p)$, $S/M$ is an extension of $C_p$ by a subgroup of order $q^\mu \mid (p-1)$;

(c) $G/S \simeq L_2(p)$, $|S/M| = p$, and $q \mid (p+1)/2$.

In all cases Theorem 6.3.3 of [2] yields $C_S(M) \subseteq M$, and so $|M| \ge p+1$. Since $Z(G) = 1$, the case that a subgroup isomorphic to $SL_2(p)$ is normal in $G$ is excluded.

(a) If $q = 2$ then $|G| \le 8p^4$, while $|S|2|L_2(p)| \ge (p+1)p^2(p^2-1)$.

Therefore, $p \le 7$. Since in this case the subgroup $O_{p'}(G)$ must be of order at most 8, we arrive at a contradiction. Case (a) is impossible.

(b) In this case $q \leq (p-1)/2$ and $G \simeq S \times L$, where $L \simeq L_2(p)$. Straightforward calculations yield the bound $|M| \leq 2p-2$. Thus, $M$ includes at most one length $p$ orbit. In particular, $M$ is an elementary abelian primary group. Hence, an arbitrary irreducible character degree $\psi$ of the group $S = M \rtimes C_p \rtimes C_d$ for $d$ dividing $p-1$ divides $dp$. Since $p^2 > |M|$, it follows that $\psi(1) \leq d$. According to [15, pp. 262–263], all irreducible character degrees of $L_2(p)$ are at most $p+1$.

By Theorem 3.7.1 of [2], every irreducible character of $G$ is the product of irreducible characters of $S$ and $L$. Therefore, $G$ lacks degree $p^2q$ characters.

(c) The group $G = S \times L$. Since $|G| \leq 2p^4q^2$, it follows that $|S||L| \leq 2p^4q^2$ and $|S| \leq 4pq^2 + 2q$. Since $L$ lacks characters of degree greater than $p+1$ (see [15, pp. 262–263]), while each irreducible character degree of $G$ is the product of an irreducible character degree of $L$ and an irreducible character degree of $S$, it follows that an irreducible character of $S$ of degree $pq$ must exist. Therefore, $|S| > p^2q^2$, whence $p < 4q$. This means that $q = (p+1)/2$.

Recall that $q^{b-2}m < 2p^2$. Since $p-1$ divides $m$ in this case, it follows that $m = t(p-1)$ for some positive integer $t$. For $b = 4$ we obtain $(p+1)^2m < 8p^2$. Then $m \leq 7$ and $p \leq 8$. Easy calculations exclude this possibility.

Suppose that $b = 3$. Then $m \leq 4(p-1)$. In this case $|M| = q^2m/(p-1) = tq^2$, where $t \leq 4$. If $t = q = 2$ then $p = 3$, which is already excluded. If $t = q = 3$ then $p = 5$. Then $p = 5$ does not divide $|Out(M)|$. In the remaining cases $S$ has an abelian Sylow $q$-subgroup. Lemma 3 implies that $S$ cannot have an irreducible character of degree $pq$.

Suppose that $b = 2$. Then $p^2 < m < 2p^2$. Furthermore, $m = t(p-1)$ for some positive integer $t$. The order of $M$ equals $qm/(p-1)$. If $t = q$ or $t = 2q$ then $S$ lacks irreducible characters of degree $pq$. In all other cases the order of $S$ is less than $p^2q^2$, and so $b \neq 2$.

Now $b = 1$ and $|G| = p^2qm$, where $p^2q < m < 2p^2q$ and $q = (p+1)/2$. In this case $|S| = pm/(p-1)$. Since an irreducible character degree must divide the group order, this implies that $S$ lacks degree $pq$ characters, but then $G$ lacks degree $p^2q$ characters.

All possibilities are excluded and the proof of Theorem 2 is complete.

## 4. Proof of Theorem 3

Theorem 2 establishes that $G$ is a $p$-solvable group. In Lemma 12, on assuming that $|G| \neq 486$, the order of the Sylow $p$-subgroup $P$ of $G$ is at most $p^3$. Lemmas 14 and 16 show that $|P| = p^2$ for $O_p(G) = 1$.

**Lemma 18.** *$O_p(G) = 1$ and $|P| = p^2$.*

Proof. Suppose that $U = O_p(G) = 1$. By Lemma 12, the subgroup $P \in \mathrm{Syl}_p(G)$ is not normal in $G$.

Lemma 3 implies that $|U| \neq p^2$. Therefore, $|U| = p$. Consequently, $H = C_G(P)$ includes the commutant of $G$, and by Lemma 7 we have $|G/H| = q^\lambda \neq 1$ and the group $G/H$ is cyclic. Lemmas 1, 2, and 6 yield

$$\Theta_H = \sum_{i=1}^{q} \chi_i,$$

where $\chi_i \in \mathrm{Irr}(P)$.

Since the order of $G/H$ is coprime to $p$ and $\Theta(1) = p^2q$, we have $\chi_i(1) = p^2$. Therefore, $H$ has $q$ conjugate irreducible characters of degree $p^2$. Since $|H| > p^4q$ and $|G/H| = q^\lambda$, it follows that $\lambda = 1$.

There is a $q$-subgroup $Q$ of $G$ not lying in $H$. It acts by conjugation on the set of degree $p^2$ irreducible characters of $H$.

The Galois group acts on the set of degree $p^2q$ characters of $G$. The total number of degree $p^2$ irreducible characters of $H$ equals $p-1$. The orbits of $Q$ contain $q$ of them. Thus, the number of degree $p^2q$ irreducible characters of $G$ is at least $(p-1)/q$. Since we have exactly one character conjugate to $\Theta$, it follows that $p-1=q$. Therefore, in this case $p=3$ and $q=2$, which is excluded in the remark above Lemma 14. Thus, $O_p(G)=1$ and $|P|=p^2$.

The proof of Lemma 18 is complete.

Theorem 2 implies that $G$ is a $p$-solvable group, while by Lemma 18 the Sylow $p$-subgroup $P$ of $G$ is an abelian group of order $p^2$. Theorem 6.3.3 of [2] yields $G=O_{p',p,p'}(G)$, and furthermore, $H=O_{p',p}(G)=M \rtimes P$, where $M=O_{p'}(G)$. Fix this notation for the rest of the proof of Theorem 3.

**Lemma 19.** *If $G \neq H$ then $|G:H|=q$ and $\Theta_H=\sum_{i=1}^{q}\chi_i$, where $\chi_i \in \mathrm{Irr}(H)$ are conjugate characters of $H$ of degree $p^2$.*

Proof. Theorem 2 shows that the $LC(\Theta)$-group $G$ is $p$-solvable and $|G|=p^2q^bm$, where $\Theta(1)=p^2q$; moreover, $(pq,m)=1$. Take $P \in \mathrm{Syl}_p(G)$ and $H=O_{p',p}(G)$. Since $P$ is an abelian group of order $p^2$, it follows that $P$ is isomorphic either to the cyclic group $C_{p^2}$ or to the elementary abelian group $C_p \times C_p$.

Therefore, $G/H \leq Out_G(P)$ is a $p'$-group. Put $f=|G:H|$. By Lemma 7, the group $G/H$ cannot have normal subgroups of prime indices distinct from $q$. Since the $p'$-group $Out_G(P)$ is isomorphic either to a cyclic group of order dividing $p-1$ (the case $P=C_{p^2}$) or to a group of order dividing $(p^2-1)(p-1)$ (the case $P \cong C_p \times C_p$), it follows that $q$ divides $p \pm 1$.

By Lemma 1 (Clifford's Theorem),

$$\Theta_H = e\sum_{i=1}^{s}\chi_i,$$

where $\chi_i$ are conjugate irreducible characters of $H$, and furthermore, $es\chi_1(1)=p^2q$, where $es$ divides $|G:H|$. In particular, $(p,es)=1$ implies that $es=q$ and $\chi_i(1)=p^2$. In any case we have a normal subgroup $H_1 \geq H$ of index $q$ in $G$. Since $\Theta_{H_1}$ is reducible by Lemma 7, it follows that $\Theta_{H_1}=e\sum_{i=1}^{q}\chi_i$, where $\chi_i(1)=p^2$. In particular, $|G:I_G(\chi_1)|=q$. Therefore,

$$\chi_1^G=\sum_{i=1}^{t}e_i\overline{\phi}_i,$$

where one of the characters, for instance $\overline{\phi}_1$, equals $\Theta$, and

$$\sum_{i=1}^{t}e_i^2=|I_G(\chi_1):H|=f/g.$$

It is clear that $H_1=I_G(\Theta)$. Since $H_1$ has $q$ irreducible characters of degree $p^2$, it follows that $|H_1| \geq (p^4q+1)$. On the other hand, the subgroup $M=O_{p'}(G)$ is not centralized by any element of $P^{\#}$ in view of Lemma 3 (Ito's Theorem). Therefore, $|M|>p^2$, and so $|H|>p^4$. Consequently, from $|G|=|G:H|<2p^4q^2$ we infer that $f=|G:H|<2q^2$. Thus, Lemma 7 leaves the following open possibilities:

(a) $|G:H|=q^2$;

(b) $|G:H| = q(q+1)$, where either $q+1$ is a power of 2 or $q+1=3$ and $q=2$;
(c) $|G:H| = q$.

But in cases (a) and (b) we have more than one irreducible character of degree $p^2q$, which is false. Hence, $|G:H| = q$ and the proof of Lemma 19 is complete.

**Lemma 20.** *If $G \neq H$ then for every irreducible character $\chi = \chi_i$ in the expansion of $\Theta_H$ there exist $p^2$ linear characters $\psi_j$ conjugate to the character $\psi = \psi_1$, so that $\chi_M = \sum_{j=1}^{p^2} \psi_j$. The group $M$ is abelian.*

Proof. Take an irreducible character $\psi$ of $M$ appearing in the expansion of the character $\chi_M$, where $\chi = \chi_i$ is an arbitrary irreducible character of $H$ defined in Lemma 19. According to Lemma 1,

$$\chi_M = e' \sum_{i=1}^{t} \psi_i,$$

where $\psi_i$ are characters conjugate to $\psi$, while $e'$ and $t$ divide $|H:M| = p^2$. In particular,

$$p^2 = \chi(1) = e't\psi(1).$$

Since $(|M|, p) = 1$, it follows that $\psi_i(1)$ is not divisible by $p$. Thus, $e't = p^2$ and $\psi_i(1) = 1$. In particular, $ker(\psi_i)$ contains the commutant of $M$. Hence, $\psi_j(y) = 1$ for every $j$ and $y \in M$. Then $\chi(y) = p^2$ for every character $\chi$ conjugate to $\chi_1$. As a result, $\Theta(y) = p^2q$. Since $\Theta$ is faithful, it follows that $y = 1$; i.e. $M' = 1$ and $M$ is an abelian group, as claimed.

**Lemma 21.** *Suppose that $G = H$ and $G = M \rtimes P$, and that $G$ lacks index $q$ normal subgroups. Then $\Theta_M = \sum_{i=1}^{p^2} \alpha_i$, where $\alpha_i \in \mathrm{Irr}(M)$ are conjugate characters of $M$ of degree $q$.*

Proof. Take an irreducible character $\alpha \in \mathrm{Irr}(M)$ of $M$ appearing in the expansion of $\Theta_M$ and consider its inertia group $I_G(\alpha)$. By Lemma 1,

$$\Theta_M = e \sum_{i=1}^{s} \alpha_i,$$

where $\alpha_i \in \mathrm{Irr}(M)$, while $e$ divides $|I_G(\alpha) : M|$ and $s$ divides $|G : I_G(\alpha)|$. Lemma 6 implies that $\Theta_M$ is a reducible character, and so $s > 1$. Since $\Theta(1) = p^2q = es\alpha(1)$, and furthermore, $(\alpha(1), p) = 1$ (the order of $M$ is not divisible by $p$), it follows that $es = p^2$ and $\alpha(1) = q$. Since $s \neq 1$, it follows that $e = 1$.

Thus, $s = p^2$. The proof of Lemma 21 is complete.

**Lemma 22.** *If $r$ is a prime divisor of $|M|$ then $M$ includes a Sylow $r$-subgroup $R$ admissible with respect to $P$.*

Proof. The group $P$ acts by conjugation on the set of Sylow $r$-subgroups of $M$. The number of them is the index of the normalizer of a Sylow subgroup, which is coprime to $|P|$. Since the length of every nontrivial $P$-orbit of the group is divisible by $p$, a Sylow $r$-subgroup admissible with respect to $P$ must exist.

The proof of Lemma 22 is complete.

**Lemma 23.** *The group $G = M \rtimes P$ includes the subgroups $QP$ and $RP$ of orders $q^bp^2$ and $r^\alpha p^2$ respectively; furthermore, $r^\alpha | m$, where $P$, $Q$, and $R$ are Sylow $p$-, $q$-, and $r$-subgroups of $G$. In particular, $M$ has Hall $\{p, q\}$- and $\{p, r\}$-subgroups.*

Proof. As we said above, $G = M \rtimes P$, where $|M| = q^bm$ and $|P| = p^2$ with $(p, qm) = 1$. Since $P$ acts by conjugation on the set of Sylow $q$-subgroups of $M$, while the number of them is not divisible by $p$, it follows that there is a Sylow $q$-subgroup $Q$ of $M$ invariant under $P$. Therefore, $G$ includes the subgroup $L = QP$ of index $m$. Lemma 22 implies that $G$ also includes the subgroup $Y = RP$ of index $\frac{q^b m}{r^\alpha}$.

The proof of Lemma 23 is complete.

**Lemma 24.** *If $G = H$ then one of the following claims holds:*
*1. $G = QC_G(P)$, where $|Q| = q^b$, and moreover, $q^b > p^2$.*
*2. $G = RC_G(P)$, where $|R| = r^\alpha$ divides $m$.*
*3. $G = C_G(a)C_G(u)$, where $P = \langle a\rangle \times \langle u\rangle$ is of order $p^2$.*

PROOF. Recall that $|G| = p^2 q^b m$ and $(p, qm) = 1$; furthermore, $q^{b-2}m < 2p^2$.

Take $R \in \mathrm{Syl}_r(G)$, where $r|m$. If $P$ acts nontrivially on $R$ then $|R| \geq p + 1$.

Suppose now that $P$ acts trivially on every Sylow subgroup of $G$ except $R$. Lemma 11 shows that $G$ has a normal subgroup $Y_1$ which is the normal closure of $P$ included in $Y = RP$. Ito's Theorem (Lemma 3) implies that $P \neq Y_1$. Hence, without loss of generality we may assume that $Y_1 = R_1 \rtimes P = \langle P^G\rangle$, where $R_1$ is a subgroup of $R$ on which $P$ acts nontrivially. In this case $G = RC_G(P)$, so that claim 2 holds.

If $P$ acts nontrivially, besides $R \in \mathrm{Syl}_r(G)$, on some $S \in \mathrm{Syl}_s(G)$, where $s|m$ and $s \neq r$, then $|S| \geq p + 1$, and furthermore, $m \geq |S||R| \geq (p + 1)(2p + 1) > 2p^2$. Hence, $m > 2p^2$. Since $q^{b-2}m < 2p^2$, it follows that $b = 1$. Therefore, $|G| = p^2 qm$, where $(m, pq) = 1$.

If for every simple $r \in \pi(M)$ the group $P$ acts trivially on some $R \in \mathrm{Syl}_r(G)$ then $G = M \times P$, which is impossible.

Suppose that for every $r \neq q$ the subgroup $P$ acts trivially on some Sylow $r$-subgroup of $M$ from $M = O_{p'}(G)$. Lemma 11 shows that $G$ has a normal subgroup $L_1$, which is the normal closure of $P$ included in $L = QP$, of order divisible by $p^2$. Lemma 3 implies that $P \neq L_1$; therefore, $L_1 = Q_1 \rtimes P = \langle P^G\rangle$, where $Q_1 \leq Q$. Then $G = QC_G(P)$; furthermore, $q^b > p^2$ and claim 1 holds.

Finally, it is possible that $P$ acts nontrivially on both $R$ and $Q$, where $|Q| = q^b$ and $|R| = r^\alpha|m$.

It is clear that $|R| \geq 1 + p$. If $|R| \geq p^2$ then $m \geq p^2 + 1$, and $mq^{b-2} \leq 2p^2$ forces $q^{b-2} \leq 1$, so that $b = 1$ or 2.

But in this case $|P| > |Q|$ and $P$ cannot act nontrivially on $Q$ with the exception of the case when $q = 2$ and $p = 3$. This case is considered above, and we assume henceforth that $b \geq 3$.

Therefore, $|R| < p^2$, and so there exists an order $p$ element acting identically on $R$. Furthermore, $m \geq |R| \geq p + 1$.

Consider $L = Q \rtimes P$. Since by Burnside's Theorem (Theorem 5.1.4. of [2]) every order $p$ element not centralizing $Q$ must act nontrivially on $Q/\Phi(Q)$, it follows that $|\Phi(Q)| \leq q$.

Therefore, the following cases are possible:

(1) $Q$ is an extraspecial group of order $q^b$, where $p$ divides $q^{b-1}-1$, and moreover, $\Phi(Q) = Z(Q) = Q'$ is of order $q$ and $|Q/\Phi(Q)| = q^{b-1}$.

In this case $b-1$ is even. Suppose that $b-1 = 2s$; then $q^{2s} - 1 = (q^s - 1)(q^s + 1)$ is divisible by $p$. Hence, $p \leq q^{\frac{b-1}{2}} + 1 < q^{b-2}$; however, $q^{b-2}m < 2p^2$. Then $m < p$, which is impossible.

(2) $Q$ is a nonelementary abelian and $|\Phi(Q)| = q$. Then $Q = \langle a\rangle \times Q_1$; furthermore, $|\langle a\rangle| = q^2$ and $|Q_1| = q^{b-2}$. It is clear that $\langle a^q\rangle$ is centralized by $P$ and

$\langle a^q \rangle \times Q_1 = \Omega(Q) \lhd L$. By Maschke's Theorem there exists $\widetilde{Q_1}$ of order $q^{b-2}$ with $\langle a^q \rangle \times \widetilde{Q_1} = \Omega(Q)$; consequently, $P$ acts trivially on $Q$, which is excluded.

(3) $Q$ is an elementary abelian of order $q^b$, and furthermore, $q^{b-1}-1 \equiv 0 (\operatorname{mod} p)$. Since $q^b = q(q^{b-1}-1) + q = qpt + q$ and $p^2 > q^b > p$, it follows that there exists an order $p$ element $a$ centralizing $Q$. Also there exists an order $p$ element $u$ centralizing $R$. Moreover, $\langle a \rangle \neq \langle u \rangle$ because $Z(G) = 1$. Thus, $C_G(a)C_G(u) \geq G$ and $P = \langle a \rangle \times \langle u \rangle$, where $a^p = u^p = 1$, i.e. claim 3 holds.

The proof of Lemma 24 is complete.

**Lemma 25.** *Let $\alpha$ be an irreducible character of the group $M$ of degree $q$ appearing in the expansion of $\Theta_M$. Then for $K \lhd M$ the character $\alpha_K$ is reducible.*

Proof. Lemmas 11 and 24 imply that $M$ is not a simple group. This is obvious if $\pi(m) = 1$. The same holds for $\pi(m) = 2$: indeed, if $G = RC_G(P)$ for some Sylow subgroup $R$ of $M$, then by Lemma 11 we have a normal subgroup of $G$ included into $M$ (cases 1 and 2 of Lemma 24). But if no nonidentity element of $P$ centralizes at least a pair of Sylow subgroups of $M$, for instance $R$ and $S$, which $P$ normalizes, and furthermore, one of them is a Sylow $r$-subgroup of $M$, while the second one is a Sylow $s$-subgroup, where $r \neq s \in \pi(m)$, then $m > (p+1)(2p+1) > 2p^2$. Hence, $b = 1$ and $|G| = p^2qm$.

In this case $P$ centralizes $Q \in \operatorname{Syl}_q(M)$. Suppose that $|\pi(m)| = 2$. Suppose that $C_P(R) = \langle a \rangle$. Then $|G : C_G(a)|$ is a power of $s$. By Lemma 11 we conclude that the subgroup $\langle a^G \rangle$ is a solvable $\{p, s\}$-subgroup, so that $M$ is not simple. If $|\pi(m)| \geq 3$ and $P^{\#}$ contains no element the index of whose centralizer is a prime power then $m \geq 2(p+1)^3$, so that $mq^{-1} > 2p^2$; this is a contradiction. Thus, in all cases $M$ is not simple.

Denote by $K$ the greatest proper normal subgroup of $M$. By Lemma 1,

$$\alpha_K = e \sum_{s=1}^{d} \lambda_s,$$

where $d$ and $e$ are positive integers dividing $|M : K|$, and $\lambda_s$ are irreducible characters of $K$ conjugate with $\lambda_1 = \lambda$. Hence, $q = \alpha(1) = ed\lambda(1)$. If $\alpha_K$ is irreducible for every $\alpha$ appearing in the expansion of $\Theta_M$ then $K$ has $p^2$ irreducible characters of degree $q$. Then its order is at least $q^2p^2 + 1$, while

$$|K| \leq |M|/(|M : K|) \leq p^2q^2 + 1;$$

this is a contradiction. Hence,

$$\alpha_K = \sum_{s=1}^{q} \lambda_s,$$

where $\lambda_s$ are linear characters.

The proof of Lemma 25 is complete.

Proof of Theorem 3. Lemma 25 implies that the intersection of the kernels of the characters $\lambda_s$ appearing in the expansion of $\Theta_K$ contains the commutant of $K$, and so it is trivial. Thus, the subgroup $K$ of $G$, of index $p^2q$ in $G$, is abelian, as claimed.

The proof of Theorem 3 is complete.

## *REFERENCES*

**1.** *Vdovin E. P.* Maximal orders of Abelian subgroups in finite simple groups // Algebra and Logic. 1999. V. 38, N 2. P. 67–83.
**2.** *Gorenstein D.* Finite Groups. New York; London: Harper and Row, Publ., 1968.
**3.** *Isaacs I. M.* Character Theory of Finite groups. New York; San Francisco; London: Acad. Press, 1976.
**4.** *Feit W.* Characters of Finite Groups. New York; Amsterdam: Yale University, 1967.
**5.** *Zenkov V. I.* The intersections of nilpotent subgroups in finite groups // Fund. Prikl. Mat. 1996. V. 2, N 1. P. 1–92.
**6.** *Kazarin L. S. and Poiseeva S. S.* On finite groups with an irreducible character large degree // Model. Anal. Inform. Sist. 2015. V. 22, N 4. P. 483–499.
**7.** *Guralnick R. M., Malle G., and Navarro G.* Self-normalizing Sylow subgroups // Proc. Amer. Math. Soc. 2003. V. 132, N 4. P. 973–979.
**8.** *Feit W.* The Representation Theory of Finite Groups. Amsterdam: North-Holland, 1982.
**9.** *Kazarin L. S.* Burnside's $p^{\alpha}$-lemma // Math. Notes. 1990. V. 48, N 2. P. 749–751.
**10.** *Camina A. R. and Camina R. D.* Implications of conjugacy class size // J. Group Theory. 1998. V. 1. P. 257–269.
**11.** *Gorenstein D.* Finite Simple Groups. An Introduction to Their Classification. New York: Plenum, 1982.
**12.** *Conway J. H., Curtis R. T., Norton S. P., Parker R. A., and Wilson R. A.* Atlas of Finite Groups. Maximal Subgroups and Ordinary Characters for Simple Groups. Oxford: Clarendon Press, 1985.
**13.** *Kazarin L. S. and Poiseeva S. S.* Finite groups with large irreducible character // Math. Notes. 2015. V. 98, N 2. P. 265–272.
14. *GAP*—Groups, Algorithms and Programming. Version 4.4.10 (Aachen, St. Andrews, 2008); http://www.gap-system.org.
**15.** *Belonogov V. A.* Representations and Characters in the Theory of Finite Groups [in Russian]. Sverdlovsk: Akad. Nauk SSSR, Ural′sk. Otdel., 1990.

*October 15, 2015*

S. S. Poiseeva
Yaroslavl′ State University, Yaroslavl′, Russia
`pss.iii@mail.ru`

*Yakutian Mathematical Journal*
*October—December, 2015. Vol. 22, No. 4*

*UDC 512.6:519.61*

# PARTICULARITIES OF INFINITE SYSTEMS

**F. M. Fedorov, O. F. Ivanova, and N. N. Pavlov**

**Abstract.** Basing on the previous results on infinite Gaussian systems, we study the key fundamental differences between general infinite systems and finite systems. In particular, we show that Fredholm's and Noether's theorems fail for general infinite systems of linear algebraic equations. In addition, we refine the row reduction method, showing in particular that it can converge though not to the solution to the infinite system under consideration. We also indicate that the row reduction method for solving homogeneous infinite systems reveals duality. The solution of homogeneous infinite systems is of a contradictory character as regards finite homogeneous systems. In particular, we show that a homogeneous infinite system can have a nontrivial solution even if its infinite determinant is nonzero.

In addition, solving a linear homogeneous infinite system necessarily reduces to a nonlinear equation, called the characteristic equation, which is impossible for finite systems.

**Keywords:** infinite Gaussian system, linear algebraic equations, Fredholm's theorems, Noether's theorem, Gauss transformation, row reduction method, homogeneous systems

## Introduction

After more than 200 years of research in infinite systems, their general theory is still absent, although particular theories are developed for isolated types of infinite systems [1, 2]. The general theory turned out exceedingly complicated on the one hand; and on the other hand, it is rather rich and widely applied in many areas of mathematics, mechanics, and physics. All that forced mathematicians all over the world to start studying particular classes of infinite systems. Active studies of infinite systems and the rise of functional analysis practically coincided in time at the end of the 19th and the beginning of the 20th centuries. The success of functional analysis and its wide applications in various areas of mathematics and physics predetermined the application of its methods to study infinite systems. Presently, about the dozen classes of infinite systems are completely understood: normal systems, regular and totally regular systems, multiplicative systems, systems with difference indices, and so on [2]. All these studies successfully applied functional analysis methods, but their application requires certain restrictive assumptions on the coefficient matrix and the right-hand side of the infinite system related to infinite systems in normed spaces. Norm convergence in Banach spaces implies strong convergence.

Thus, all available works [2, 3] admit that the right-hand sides $b_j$ are jointly bounded: $b_j \leq B$. In addition, the weakest restriction on the coefficients of the system $a_{i,j}$ which must be imposed amounts to $\sum_{i=1}^{\infty} |d_{i,j}| < \infty$, where $a_{i,i} = 1 + d_{i,i}$ for $i = j$ and $a_{i,j} = d_{i,j}$ for $i \neq j$. The systems for which these conditions fail a priori

The authors were supported by the Ministry of Science and Education of the Russian Federation (Grant No. 3047).

cannot be considered as objects of study, which has eventually led to a critical situation.

Precisely the development of the theory of periodic infinite systems [1] made it possible to study infinite systems from general viewpoints and enabled us in the recent years to push ahead from the crisis.

In all our studies we stick to the concept of weak convergence; more exactly, the coordinate-wise convergence. Thus, right away we avoid restrictive assumptions on the coefficients and right-hand sides of an infinite system, sticking only to the definition of solution to a system.

In this article, basing on the previous results and examples of solutions to concrete infinite systems, we exhibit the main distinctive features of their solution as compared to solutions of finite systems. In this regard, we also indicate the difficulties that might happen while solving them.

For the main facts on infinite systems, matrices, determinants, and minors, see [1–3].

First of all, let us dwell on the well-known theorems of Fredholm and Noether on the solutions to finite systems of linear equations.

## 1. Fredholm's and Noether's Theorems

To begin with, consider the question of solvability of a finite system of $n$ linear algebraic equations with $n$ unknowns. We can express every system as one linear equation

$$Ax = b, \tag{1}$$

where $b$ is a given vector, $x$ is the required solution vector, and $A$ is a linear operator in the $n$-dimensional Euclidean space $\mathbb{R}^n$ defined by the matrix of the system.

Recall the main facts of linear algebra concerning this finite system [4, 5].

1. Eq. (1) is solvable for every right-hand side if and only if the corresponding homogeneous equation $Ax = 0$ admits only the trivial solution $x = 0$.

2. Eq. (1) is solvable for every right-hand side if and only if the adjoint equation

$$A^*y = f \tag{2}$$

is solvable for every right-hand side.

3. The equations $Ax = 0$ and $A^*y = 0$ have the same number of linearly independent solutions.

4. If the homogeneous equation $Ax = 0$ admits a nontrivial solution then the inhomogeneous equation (1) is solvable if and only if the right-hand side $b$ is orthogonal to all solutions of the adjoint homogeneous equation $A^*y = 0$.

Properties 1, 3, and 4 for finite systems of algebraic equations are reflected in Fredholm's theorems.

We should emphasize that properties 1 and 2 hold when the determinant $|A|$ of the system (1) in nonzero, while property 4, in contrast, is interesting principally when the determinant $|A|$ is zero.

Note now Noether's theorems for finite systems [6]: *the difference between the number $n$ of linearly independent solutions to the homogeneous equation $Ax = 0$ and the number $n'$ of linearly independent solutions to the adjoint equation $A^*y = 0$ equals the index $\varkappa$ of the operator $A$ (the index of (1)): $n - n' = \varkappa(A)$.*

What is the situation in the case of infinite systems? Which properties hold and which do not in the theory of infinite systems? Do some new properties appear? We can answer these questions only basing on our results. Therefore, consider a concrete example of an infinite system. We should emphasize that it arose in mathematical physics while solving a boundary value problem by the boundary method [7, 8], i.e., it is not just some abstract system.

EXAMPLE 1. Consider the following basic model of heat problem with variable boundary conditions [7]:

$$\begin{gathered}
\frac{\partial T(x,t)}{\partial t} = \alpha \frac{\partial^2 T(x,t)}{\partial x^2}, \quad 0 < x < 1, \\
\frac{\partial T(x,t)}{\partial x} = 0, \quad x = 0, \\
T(x,t) = V \exp(\nu t), \quad x = 1, \\
T(x,t) = 0, \quad t = 0.
\end{gathered} \tag{3}$$

The first stage in solving (3) by the boundary method [8] amounts to solving the infinite system [7]

$$\sum_{p=0}^{\infty} \frac{(2j+2p)!}{(2p)!} x_{j+p} = V \exp(\nu t) \left( \frac{\nu}{\alpha} \right)^j, \quad j = \overline{0, \infty}, \ b = \frac{\nu}{\alpha} = \text{const} > 0, \tag{4}$$

where $x_j = a_j(t)$ and $a_j(t)$ are the coefficients of the power series expansion of the temperature function $T(x,t)$ in the spatial coordinate $x$.

Clearly, all row series of absolute values of the matrix entries of (4) diverge:

$$\sum_{p=0}^{\infty} |a_{j,j+p}| = \sum_{p=0}^{\infty} \frac{(2j+2p)!}{(2p)!} = \infty, \quad j = 0, 1, 2, \ldots,$$

In addition, the right-hand sides of (4) for $b > 1$ are not jointly bounded. Nevertheless, this system has a solution obtained using the row reduction method in the broad sense [7]:

$$\begin{aligned}
x_j^{(k)} = {} & \frac{(-1)^{j+1} V \exp(\nu t)}{(2j)! \cosh \sqrt{\frac{\nu}{\alpha}}} \left\{ \left[ \frac{\pi(2k+1)}{2} \right]^{2j} - \left( -\frac{\nu}{\alpha} \right)^j \right\} \\
& + \frac{(-1)^j x_0}{(2j)!} \left[ \frac{\pi(2k+1)}{2} \right]^{2(j-1)}, \quad k, j = 0, 1, 2, \ldots,
\end{aligned} \tag{5}$$

where $x_0$ is an arbitrary real number.

Since $x_0$ is arbitrary, we can express the solution to (5) in general form as

$$\begin{aligned}
x_j^{(k)} = {} & \frac{V \exp(\nu t)}{(2j)! \cosh \sqrt{\frac{\nu}{\alpha}}} \left( \frac{\nu}{\alpha} \right)^j \\
& + \frac{(-1)^j x_0}{(2j)!} \left[ \frac{\pi(2k+1)}{2} \right]^{2j} = x_j + \tilde{x}_j^{(k)}, \quad k, j = 0, 1, 2, \ldots,
\end{aligned} \tag{6}$$

where $x_j$ is a particular solution to the inhomogeneous system (4) and $\tilde{x}_j^{(k)}$ are solutions to the homogeneous system (4) independent with respect to $k$, i.e., for $b = 0$.

It is clear from (6) that for infinite systems in general properties 1–4 of finite linear equations (1) are violated; even more so, Fredholm's theorems fail, and therefore Fredholm's alternative as well. In addition, Noether's theorem breaks down. Indeed, $\tilde{x}_j^{(k)}$ is a nontrivial solution to the corresponding homogeneous system (4), even though the infinite inhomogeneous system (4) has a solution, i.e., property 1 fails.

As for property 1, we can give more general arguments. Express an infinite system also as a linear equation (1), but in matrix form $AX = B$. Assume that the infinite determinant $|A|$ exists and is nonzero. However, the existence (or not) of the inverse matrix $A^{-1}$ in general is independent of the presence of nontrivial solutions to the corresponding homogeneous equation. Thus, the solvability of an infinite system is also independent of the existence of nontrivial solutions to the homogeneous system. Assume now that at least a left inverse matrix $A^{-1}$ exists. Then (1) yields $X = A^{-1}B$, but the product $A^{-1}B$ need not exist because the corresponding series may diverge, and then solutions to the system need not exist.

Property 2 also fails. For a Gaussian matrix $A$ the transpose $A^T$, which in our case is the same as the adjoint matrix $A^*$, is a triangular infinite matrix. Cook showed [9] that these matrices have unique inverse matrices. This implies that the adjoint equation (2) has a unique solution for each right-hand side $f$ because triangular matrices have finite rows and so they can be multiplied:

$$(A^*)^{-1}A^*y = (A^*)^{-1}f, \quad y = (A^*)^{-1}f. \tag{7}$$

However, (1) is not always solvable even if the infinite determinant is nonzero; we discuss that in detail below in Examples 2 and 3. As (7) implies, the adjoint homogeneous equation (2) always admits only the trivial solution; consequently, as Example 1 shows, property 3 fails too. Since the solvability of (1) is independent of the solvability of the homogeneous adjoint equation (2), it becomes evident that (4) fails. The solution (6) in Example 1 points out directly at a violation of Noether's theorem.

Therefore, for general infinite systems of linear algebraic equations (1) we lose the solvability theorems, Fredholm's and Noether's theorems, and it is impossible to study (1) using Fredholm or Noether operators $A$.

The main difference of finite systems from infinite ones is that even if the infinite determinant is nonzero the homogeneous infinite system can have nontrivial solutions, which is impossible for finite systems.

However, certain infinite systems reduce to finite systems, and the prefix "quasi" was suggested for them.

## 2. Pseudoinfinite Systems

Along with the prefix "quasi", introduce the concept of pseudoinfinite systems. Apparently, this term was applied for the first time to regular infinite systems [3]. Call *quasiregular* every system in which the regularity condition is satisfied only in all rows starting at some, i.e., the regularity condition is violated for finitely many equations. As [3] showed, the existence of a solution to these system reduces to that for a finite system. By analogy, we can introduce the concept of quasihomogeneous systems. Quasiperiodic systems appeared in the same fashion [2]. Here we introduce the concept of pseudoinfinite systems, but of a different kind than the above quasisystems. There exist infinite systems taking an intermediate position between finite and infinite systems. This is best seen by solving a concrete infinite system.

Example 2. Consider the infinite system with difference indices:

$$\sum_{p=0}^{\infty} a_p x_{j+p} = b^j, \quad j = 0, 1, \dots, \tag{8}$$

where $a_0 = 1$, $a_1 = a$, $a_2 = a_3 = \dots = 0$, $a = \text{const}$, $b = \text{const}$, and $1 + ab \neq 0$. Find a strictly particular solution to (8), as soon as it exists.

Solution. Observe that the Gaussian system (8) differs substantially from the general infinite Gaussian system; namely, each equation involves only finitely many terms (unknowns). In this case only two terms are present. For this reason, two fundamentally different approaches arise when we apply the row reduction method. Firstly, we can solve (8) without throwing terms away, but restricting only the number of equations. Then the finite system has degenerate matrix. Therefore, in this case it is convenient to use the row reduction method in the broad sense [1, 2]. Secondly, we can close this finite system by throwing away one term just in the last equation and putting $x_n = b^n$. Then it becomes possible to use the row reduction method in the narrow sense [1, 2], namely, the usual simple reduction.

The infinite system (8) is also attractive in that we can solve it in at least four different ways. Its general solution is given in [1] as an example of solving periodic infinite systems:

$$x_j = \frac{b^j}{1+ab} + \frac{(-1)^j C}{a^j}, \quad 1 + ab \neq 0, \quad j = 0, 1, \dots, \tag{9}$$

where $C$ is an arbitrary constant.

Secondly, we can consider (8) as a difference equation. The theory of those appeared in the classical works, for instance in [10]; naturally, applying it, we obtain the solution (9).

The coefficient matrix $A$ of (8) is

$$A = \begin{pmatrix} 1 & a & 0 & 0 & . & 0 & 0 & . \\ 0 & 1 & a & 0 & . & 0 & 0 & . \\ 0 & 0 & 1 & a & . & 0 & 0 & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & 1 & a & . \\ 0 & 0 & 0 & 0 & . & 0 & 1 & . \\ . & . & . & . & . & . & . & . \end{pmatrix}, \tag{10}$$

and its infinite determinant obviously equals 1; i.e., $|A| = 1$.

Below we solve (8) in two more ways. Firstly, solve (8) using the general theory of infinite systems [11], according to which a strictly particular solution to an infinite Gaussian system, whenever it exists, is expressed [11] as

$$x_j = \Delta^{(j+1)} = \sum_{p=0}^{\infty} (-1)^p A_p(j) b_{j+p}, \quad j = 0, \dots, \infty. \tag{11}$$

Furthermore, $A_p(j)$ is defined recursively as

$$A_p(j) = \sum_{k=0}^{p-1} (-1)^{p-1-k} a_{j+k,j+p} A_k(j), \quad A_0(j) = 1, \tag{12}$$

where $a_{j+k,j+p}$ are the matrix entries of (10) and $\Delta^{(j+1)}$ is the Cramer determinant, which is the determinant $|A|$ with column $j+1$ replaced by the column of right-hand sides $b_k = b^k$ of (8).

Actually, $A_p(j)$ constitute the sequence of principal minors of the determinant

$$A_n(j) = \begin{vmatrix} a_{j,j+1} & 1 & \dots & 0 & 0 \\ a_{j,j+2} & a_{j+1,j+2} & \dots & 0 & 0 \\ . & . & \dots & . & . \\ . & . & \dots & . & . \\ a_{j,j+n-1} & a_{j+1,j+n-1} & \dots & a_{j+n-2,j+n-1} & 1 \\ a_{j,j+n} & a_{j+1,j+n} & \dots & a_{j+n-2,j+n} & a_{j+n-1,j+n} \end{vmatrix}; \tag{13}$$

we put $A_0(j) = 1$, $A_1(j) = a_{j,j+1}$, $A_2(j) = \begin{vmatrix} a_{j,j+1} & 1 \\ a_{j,j+2} & a_{j+1,j+2} \end{vmatrix}$, and so on. In our case $A_p(j)$ are the order $p$ principal minors of the determinant of the matrix

$$\begin{pmatrix} a & 1 & 0 & 0 & . & 0 & 0 & . \\ 0 & a & 1 & 0 & . & 0 & 0 & . \\ 0 & 0 & a & 1 & . & 0 & 0 & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & a & 1 & . \\ 0 & 0 & 0 & 0 & . & 0 & a & . \\ . & . & . & . & . & . & . & . \end{pmatrix}. \tag{14}$$

Calculating this principal minors or using (12), we easily obtain $A_p = a^p$. Then (11) yields

$$\Delta^{(j+1)} = b^j \sum_{p=0}^{\infty} (-1)^p a^p b^p, \quad j = 0, \dots, \infty. \tag{15}$$

However, this series converges if and only if $|ab| < 1$, and in this case we find

$$x_j = \Delta^{(j+1)} = \frac{b^j}{1 + ab}, \quad 1 + ab \neq 0. \tag{16}$$

Inserting this expression into (8), we verify that it is a strictly particular solution to (8) for $|ab| < 1$. Therefore, the condition $|ab| < 1$ is necessary and sufficient for the convergence of the row reduction method. However, it is obvious that this condition is simultaneously a condition on both the matrix $A$ and the right-hand side $b$ of (8). Meanwhile, it is known [12] that the linear operator defined by the coefficient matrix $A$ of the infinite system must be expressible as the sum of a positive definite operator and a totally continuous operator. Under these conditions $A$ admits reduction with respect to all orthonormal bases; hence, we obtain a strong condition for the convergence of reduction. This result is a corollary of strong convergence, which is convergence in the norm of the space, while we consider weak convergence, which is coordinatewise convergence, which yields a weaker condition for the convergence of the row reduction method.

It is obvious that the solution of (9) is independent of $|ab|$ being greater or less than 1. Together with that, we have the main theorem [11]: *an infinite Gaussian system is consistent if and only if it admits a strictly particular solution.* As we showed above, (8) with $|ab| > 1$ lacks strictly particular solutions; consequently, the main theorem shows that it is inconsistent, although (8) has solution (9) for $|ab| > 1$ as well. Seemingly this is a would-be contradiction, so what is the matter? Applying the row reduction method to prove the theorem mentioned, we assume that in each equation (more exactly, in infinitely many equations) we throw away infinitely many terms. In other words, we consider the infinite Gaussian system in full form in the sense that infinitely many equations involve infinitely many unknowns, while (8), as indicated above, in each equation involves just two unknowns, which makes it possible to obtain a solution without reduction. This is our fourth approach. Thus, consider (8) as one recurrence

$$x_{j-1} + ax_j = b^{j-1}, \quad j = 1, 2, \dots. \tag{17}$$

Solving this for $x_j$, we find

$$x_j = \frac{b^{j-1} - x_{j-1}}{a}. \tag{18}$$

Repeating this formula $j$ times, we obtain

$$x_j = \frac{1}{a^j}\left(\sum_{p=0}^{j-1}(-1)^p(ab)^{j-p-1} + (-1)^j x_0\right), \quad j > 0, \tag{19}$$

where $x_0$ is an arbitrary constant.

Express $x_0$ as $x_0 = \frac{1}{1+ab} + C$, where $C$ is an arbitrary constant. Insert the last expression into (19) and make simple rearrangements. This yields

$$x_j = \frac{1}{a^j}\left(\sum_{p=0}^{j-1}(-1)^p(ab)^{j-p-1} + \frac{(-1)^j}{1+ab} + (-1)^j C\right) = \frac{b^j}{1+ab} + \frac{(-1)^j C}{a^j}, \quad j > 0. \tag{20}$$

Thus, we arrive at the solution (9).

Since $\frac{(-1)^j C}{a^j(1+ab)}$ is a general solution to the corresponding homogeneous system, it is obvious that the expression

$$\frac{1}{a^j}\left(\sum_{p=0}^{j-1}(-1)^p(ab)^{j-p-1}\right), \quad j > 0,$$

is a particular solution to (8) obtained as a consequence of the finiteness of each equation of (8). Together with that, for $|ab| > 1$ it is not a strictly particular solution, although that is so for $|ab| < 1$, as we showed above. Therefore, the infinite system (8) behaves partly like an infinite system and partly like a finite system. Naturally, in the latter case some properties of infinite systems are lost; in particular, the row reduction method diverges because (19) involves only a finite sum, while the series, like in (15), is absent.

Now solve the infinite system (8) using the row reduction method in the narrow sense, i.e., apply simple reduction. In this case, as we said above, by the finiteness of the equations, throwing away only the last term of the last equation, we obtain the system of three equations

$$\begin{cases} x_{n-2} + ax_{n-1} = b^{n-2}, \\ x_{n-1} + ax_n = b^{n-1}, \\ x_n = b^n \end{cases} \tag{21}$$

with three unknowns $x_{n-2}$, $x_{n-1}$, $x_n$. It becomes clear that for arbitrary $n \geq 2$ all equations of (21) correspond precisely to the equations of the infinite system (8) except for the last. Solving this system, we obtain $x_{n-2} = b^{n-2} - ab^{n-1} + a^2b^n$, $x_{n-1} = b^{n-1} - ab^n$, and $x_n = b^n$. Hence, induction yields

$$x_{n-j} = \sum_{p=0}^{j} (-1)^p a^p b^{n-j+p}, \quad j = 0, \dots, n.$$

Reindexing, we finally obtain the reduced solution

$$\overset{n}{x}_j = \sum_{p=0}^{n-j} (-1)^p a^p b^{j+p}.$$

Passing to the limit as $n \to \infty$ in the last expression, we arrive at the series (15) for $|ab| < 1$, which confirms the necessity of applying the row reduction method in the narrow sense while solving inhomogeneous infinite systems.

Thus, infinite systems with infinitely many finite equations do not fully have the properties of general infinite systems, but share some features of finite systems. Hence, we can give them a special name, for instance, semi-infinite or quasi-infinite systems, and study them separately from general (full) systems. We suggest to apply the term *pseudoinfinite systems* and study them separately from general (full) systems. A typical example of pseudoinfinite system is an infinite system with a diagonal matrix: the infinite system $a_j x_j = b_j$, where $a_j \neq 0$. Actually, this system is finite, as it is one equation with one unknown. It is obvious that its solution is $x_j = \frac{b_j}{a_j}$. However, when we treat it as infinite, the results are clearly absurd already because for $|a| > 1$ the infinite determinant of this system does not exist. Therefore, by an infinite system in general we understand a system containing infinitely many equations with infinitely many nonzero coefficients ($a_{j,i} \neq 0$).

## 3. The Nonlinear Character of Solutions to Infinite Homogeneous Systems. Duality of the Row Reduction Method

In order to demonstrate the nonlinear character of solutions to homogeneous infinite systems, return to Example 1, more precisely, to solution (6) to (4) in the homogeneous case ($b = 0$):

$$x_j^{(k)} = \frac{x_0}{(2j)!} \left[ \frac{\pi(2k+1)}{2} \right]^{2j}, \quad k = 0, 1, 2, \dots, \; j = 1, 2, 3, \dots, \tag{22}$$

where $x_0$ is an arbitrary real number.

We emphasize that for $x_0 = 1$ this expression yields a basis of an infinite-dimensional subspace of nontrivial solutions to the homogeneous infinite system (4), while the solution of type (22) itself for $x_0 \equiv x_0^k = \text{const}$ is called in [1] the *fundamental solution* to the homogeneous system.

Let us repeat the method for obtaining (22) in more details to answer the question: What is so special about it? To this end, consider the homogeneous periodic Gaussian system

$$\sum_{p=0}^{\infty} a_{j,j+p} x_{j+p} = 0, \quad j = 0, 1, 2, \dots, \tag{23}$$

where the coefficients $a_{j,j+p}$ satisfy $a_{j,j+p} = a_p a_{j+p,j+p}$ by the periodicity of this system, and apply to it the general approach of [11] to studying infinite systems.

In order to solve (23), firstly we shorten it using the row reduction method in the broad sense and obtain a solution to the cutoff homogeneous finite Gaussian system in the form [11]:

$$\overset{n}{x}_j = -S_{n-j}\overset{n}{x}_{j+1}, \tag{24}$$

where

$$S_{n-j} = \frac{a_{j,j+1}}{a_{j,j}} + \sum_{p=2}^{n-j} \frac{(-1)^{p+1} a_{j,j+p}}{a_{j,j}\prod\limits_{k=1}^{p-1} S_{n-j-k}}, \quad S_1 = \frac{a_{n-1,n}}{a_{n-1,n-1}}, \quad j = \overline{1, n-2}. \tag{25}$$

Here $S_{n-j}$ is a function of $j$, i.e., $S_{n-j} = S_{n-j}(j)$.

Previously we showed [1, 11] that if the limit $\lim\limits_{n\to\infty} S_{n-j}(j) = S(j)$ exists and we can pass to the limit in (25) termwise then $x_j = \lim\limits_{n\to\infty} \overset{n}{x}_j$. This enables us to find a nontrivial solution to (23). Furthermore, the numbers $S(j)$ amount to the so-called characteristic values of the corresponding nontrivial solution to (23). Solving the arising recurrence, we obtain the nontrivial solution

$$x_j = \frac{(-1)^j x_0}{\prod\limits_{k=0}^{j-1} S(k)}, \quad j = \overline{1, \infty}, \tag{26}$$

where $x_0$ is an arbitrary real number, and $S(j)$ are unknown characteristic values.

Passing to the limit in (25) yields the system of nonlinear equations

$$\sum_{p=0}^{\infty} \frac{(-1)^p a_{j,j+p}}{a_{j,j}\prod\limits_{k=0}^{p-1} S(j+k)} = 0, \quad j = 0, 1, 2, \dots, \tag{27}$$

for determining $S(j)$, where we put $\prod_{k=0}^{-1} S(j+k) = 1$ for every $j$ to unify notation.

We should note that, inserting (26) into (23), we also obtain (27) because $x_0$ is arbitrary. If (27) lacks solutions for $S(j)$ then the homogeneous system (23) admits only the trivial solution.

We should particularly emphasize the following. In fact the applicability of the row reduction method to homogeneous infinite systems ends with (26) and (27). To elucidate this point, return to Example 1, or more precisely, to (22).

Since the solutions to systems with periodic and difference indices are isomorphic, it suffices to consider a system with difference indices ($a_{j,j} \equiv 1$) [1]. The isomorphism goes via the correspondence $y_j = \frac{x_j}{a_{j,j}}$, where $y_j$ is a solution to the periodic system, while $x_j$, to the system with difference indices. Then instead of (23) we can consider the system ($a_p = \frac{1}{(2p)!}$):

$$\sum_{p=0}^{\infty} \frac{1}{(2p)!} x_{j+p} = 0, \quad j = 1, 2, 3, \dots. \tag{28}$$

First of all we are interested in the characteristic values $S(j)$ that determine independent nontrivial solutions to homogeneous systems. This question is studied in detail in [13]. Basing on it, we may assume that $S(j) = S = \text{const}$; i.e., the characteristic values are independent of the index $j$, as the structure of (28) itself indicates. Then, (26) obviously initiates the solution

$$x_j = \frac{(-1)^j x_0}{S^j}, \quad j = 1, 2, \dots, \tag{29}$$

where $x_0$ is an arbitrary real number. Furthermore, we determine $S$ from the nonlinear equation resulting from (27) expressed as one equation:

$$\sum_{p=0}^{\infty} \frac{(-1)^p}{(2p)!S^p} = 0. \tag{30}$$

It is obvious that, basing on this relation, we can define the analytic function

$$f(x) = \sum_{p=0}^{\infty} \frac{(-1)^p}{(2p)!} x^p \tag{31}$$

and find its zeroes. In [1] this function is called the *characteristic* of periodic systems themselves, and of (4) in particular.

Let us make a digression. For finite systems, applying somehow the row reduction method in the broad sense, we obtain the corresponding solution similar to (24), (25), but with the difference that all parameters involved in it are available, while $S$ in (29) is an unknown which we should determine from the nonlinear equation (30). This is the main difference between solving finite and infinite homogeneous systems, because the new unknown $S = S(j) = \lim_{n\to\infty} S_{n-j}$ appears in the infinite case.

In general, the method for solving (31) is obviously independent of the method for solving the infinite system (28). Consequently, reflecting the linear character of solution to the homogeneous infinite system (28), the applicability of the row reduction method ends with (29) and (31). Therefore, the row reduction method in the broad sense yields only the structure of a nontrivial solution, when it exists, and a relation for determining an unknown, which is $S$ in this case. Let is find the value of $S$ in the case of characteristic $f(x)$. Obviously, (31) for the characteristic $f(x)$ yields $f(x) = \cos\sqrt{x} = 0$. From this we determine all zeroes of the function $f(x)$ as $x'_k = \left[\frac{\pi(2k+1)}{2}\right]^2$. Consequently, $S_k = \frac{1}{x'_k} = \left[\frac{2}{\pi(2k+1)}\right]^2$. Therefore, taking the periodicity of (4) into account, we arrive at the solution (22).

We emphasize once again that the row reduction method always converges only when a nontrivial solution to (23) exists. This assertion is quite simple to interpret. Suppose that a set $\{x_j\}_0^\infty$ of numbers constitutes a nontrivial solution to (23). Make the substitution $\frac{x_j}{x_{j+1}} = -S(j)$. The structure of (23) enables us to assume that $S(j) = S = \text{const}$, i.e., it is independent of the index $j$. Then we have $x_j = -Sx_{j+1}$. Solving this recurrence, we obviously obtain (29). The rest is already known.

However, by chance, the process (25) under certain conditions acts as a numerical algorithm for finding a concrete root of (30); i.e., it reflects the nonlinear character of solution of (23). In this role the row reduction method does not always converge, and here we see the duality of the method caused by the duality of solution of (23). The convergence of the process (25) for finding the zeroes of algebraic equations of infinite order is studied in more detail in [1, 14, 15]. These articles show that if the equation (30) has the unique solution with the smallest absolute value then the process (30) converges to its solution. If there are several zeroes then (25) does not converge. It is worth noting that (25) generalizes the classical Bernoulli method for calculating the zero of a polynomial with the greatest absolute value [15].

## 4. Difficulties in Searching for Solutions to Infinite Systems

Here we note only those difficulties in solving infinite systems which arise due to principal differences between finite and infinite systems.

Consider an infinite system given in operator form (a) or matrix form (b):

$$\text{(a)} \quad Ax = b, \quad \text{(b)} \quad AX = B, \tag{32}$$

where $A$ is a linear operator in matrix form or the matrix itself, $x$ and $X$ are unknown columns, $b$ and $B$ are columns of right-hand sides.

Therefore, we often treat an infinite system of equations as one linear equation (32(a)) and solve it using the theory of linear operators. However, this approach cannot fully solve infinite systems of equations, as strong arguments indicate.

Firstly, as we showed above, Fredholm's theorem and Noether's theorem, valid for finite systems, fail for infinite systems. Considering an infinite matrix as a linear operator, we obtain only a strictly particular solution to the infinite system, which in the case of a homogeneous system is only the trivial solution. Together with that, a detailed solution in the homogeneous case of the system in Example 1 shows that the matrix $A$, i.e., the operator $A$, in fact plays no role in obtaining nontrivial solutions to the homogeneous system. Therefore, no conditions imposed on $A$ beforehand guarantee the uniqueness of solution.

Secondly, the following difference of infinite systems from finite ones is important for studying the former. Consider an infinite system (32 (b)) expressed in matrix form $AX = B$, where $A$ is a Gaussian matrix, which moreover has a unique inverse matrix $A^{-1}$. Then the solution to the system in the finite case is $X = A^{-1}B$. However, this solution, always valid for finite systems, is not always a solution to an infinite system. Consider this assertion on a concrete example.

Take a general infinite system in the Gaussian form ($a_{j,j} \neq 0$):

$$\sum_{p=0}^{\infty} a_{j,j+p} x_{j+p} = b_j, \quad j = 1, 2, 3, \dots. \tag{33}$$

Without loss of generality we may assume that the Gaussian coefficient matrix $A$ of (33) has nonzero infinite determinant $|A|$. Since Gaussian matrices satisfy $a_{j,j} \neq 0$, formally dividing by the diagonal entries, we obtain a system with diagonal entries equal to 1, and so nonzero determinant.

Then [11]

$$x_j = \sum_{p=0}^{\infty} (-1)^p \frac{A_p(j) b_{j+p}}{a_{j+p,j+p}} + \frac{(-1)^j x_0}{\prod\limits_{k=0}^{j-1} S(k)}, \quad j = 1, 2, \dots, \tag{34}$$

is a particular solution to the inhomogeneous Gaussian system (33) similar to the solution (6) and accounting for a nontrivial solution, if it exists, to the homogeneous system (33) (with $b_j = 0$ for all $j$).

Clearly, the series on the right-hand side is a strictly particular solution to (33); moreover, on the one hand, it equals the infinite Cramer determinant $\Delta^{(j)}$, and on the other hand, the value of the matrix row of $A^{-1}B$. The latter follows since the infinite inverse matrix $A^{-1}$ for every Gaussian system (33) is of the form

$$A^{-1}=\begin{pmatrix} 1 & -A_1(1) & A_2(1) & . & (-1)^p A_p(1) & (-1)^{p+1}A_{p+1}(1) & . \\ 0 & 1 & -A_1(2) & . & (-1)^{p-1}A_{p-1}(2) & (-1)^p A_p(2) & . \\ 0 & 0 & 1 & . & (-1)^{p-2}A_{p-2}(3) & (-1)^{p-1}A_{p-1}(3) & . \\ . & . & . & . & . & . & . \\ . & . & . & . & (-1)^{p+1-j}A_{p+1-j}(j) & (-1)^{p+2-j}A_{p+2-j}(j) & . \\ . & . & . & . & . & . & . \\ . & . & . & . & -A_1(p) & A_2(p) & . \\ 0 & 0 & 0 & . & 1 & -A_1(p+1) & . \\ 0 & 0 & 0 & . & 0 & 1 & . \\ . & . & . & . & . & . & . \end{pmatrix}. \tag{35}$$

It is easy to obtain this expression from (11)–(13).

Return now to Example 2 and take the values $a = -10$ and $b = 1$ of the parameters. Then the coefficients of $A^{-1}$ for the matrix (10) by (35) are

$$a'_{i,j} = \begin{cases} 10^{j-i}, & j \geq i \\ 0, & j < i \end{cases}.$$

We can also verify this by the straightforward multiplication of $A^{-1}$ and $A$. The formal product $A^{-1}B$ yields the column with entries $\bar{b}_j = \sum\limits_{p=0}^{\infty} 10^p b_{j+p}$ for $j \geq 1$. However, these series need not converge, for instance, for $b_j \equiv 1$. Then the multiplication of infinite matrices is impossible by definition; in addition, the infinite Cramer determinant $\Delta^{(j)}$ does not exist. Consequently, in this case the row reduction method fails to converge and the question of solving the pseudoinfinite system (8) remains open. Observe that for a full infinite system (each equation of the system involves infinitely many unknowns) this situation demonstrates that the system under consideration is not solvable [11]. Suppose now that $b = 1/20$. Then the last series converges, the matrix multiplication works and, as we showed above, the solution to (8) is of the form (16) and equals the Cramer determinant. For $b = 1$ the matrix multiplication is impossible; nevertheless, (8) admits the solution (16), but it is not equal to the Cramer determinant because the latter fails to exist.

Although the row reduction method can converge to the Cramer determinant, the latter need not satisfy this system; i.e., in this case Cramer's rule fails.

Example 3. Consider the infinite Gaussian system

$$\sum_{p=0}^{\infty} x_{j+p} = \frac{b^j}{(1-b)^2}, \quad b \neq 1,\ j = 0, 1, 2, \ldots. \tag{36}$$

Find a strictly particular solution to (36) using (11) and (12). Calculate the values of $A_p(j)$ from the recurrence (12). Then $A_1(j) = a_{j,j}A_0(j) = 1$ and $A_2(j) = -A_0(j) + A_1(j) = 0$; consequently, $A_p = 0$ for $p \geq 2$. Taking these values of $A_p(j)$, we find from (11) that

$$\Delta^{(j+1)} = \sum_{p=0}^{\infty} (-1)^p A_p(j) b_{j+p} = b_j - b_{j+1} = \frac{b^j}{1-b} < \infty$$

for every fixed $j$ and arbitrary $b \neq 1$; hence, the row reduction method converges, and the Cramer determinant $\Delta^{(j+1)}$ exists independently of $j$.

Let us find the condition for $\Delta^{(j+1)}$ to be a solution to (36), i.e., a sufficient condition for the existence of a strictly particular solution.

Inserting $x_j = \Delta^{(j+1)} = \frac{b^j}{1-b}$ into (36) yields

$$\sum_{p=0}^{\infty} \Delta^{(j+1+p)} = \frac{b^j}{(1-b)} \sum_{p=0}^{\infty} b^p, \tag{37}$$

but the latter series converges if and only if $|b| < 1$. Consequently, for $|b| < 1$ the system (36) is satisfied, and so $x_j = \frac{b^j}{1-b}$ is a strictly particular solution.

It is obvious that for $|b| \geq 1$ the series in the right-hand side of (37) does not converge; consequently, no strictly particular solution exists, and so (36) is inconsistent by the main theorem. Hence, although the infinite determinant $\Delta^{(j+1)}$ exists, i.e., the row reduction method converges, it cannot be a solution to the system. In this case Cramer's rule fails to yield a solution to the infinite system.

Let us also pay attention to another aspect. As observed in [16], an infinite matrix can have infinitely many inverse matrices, which, obviously, leads to additional difficulties in solving these systems.

Thirdly, solutions to the homogeneous system are independent of the structure of the infinite determinant of the system. Therefore, if the infinite determinant is nonzero then the homogeneous system can have a nontrivial solution, which is impossible for finite systems. The subspace of solutions to a homogeneous system can even be infinite-dimensional. This circumstance creates additional difficulties in studying the uniqueness of solutions to infinite systems.

Fourthly, the nonlinear character of solutions to infinite homogeneous systems comes out only when we solve them by the row reduction method in the broad sense, i.e., in every cutoff system the number of unknowns is greater by one than the number of equations. More exactly, we assume that the finite cutoff system for each $n$ has degenerate matrix. The row reduction method in the broad sense always converges provided that there exists a nontrivial solution to the homogeneous system. Furthermore, it reduces the infinite system to a nonlinear (characteristic) equation. However, under certain conditions it can act as a numerical algorithm for finding a concrete root of the characteristic equation of the homogeneous system, and thereby reflect the nonlinear character of its solution. In this role the row reduction method in the broad sense does not always converge, and this reveals the duality of this method caused by the duality of solution of the homogeneous system.

Above all, note the contradictory character of solution (34) as compared to the theory of finite systems in two aspects. Firstly, if the homogeneous system (33) has a nontrivial solution then its solutions $\{y_j\}$ can be linearly dependent, which contradicts the notions of the theory of finite systems because the determinant of (33) is nonzero. Secondly, the solution of a linear homogeneous system necessarily reduces, as we have seen above, to a nonlinear equation, which cannot happen for finite systems.

In addition, as we already noted in Section 1, the operator $A$ represented by the infinite matrix $A$ of (33) is neither a Fredholm operator nor a Noether operator; thus, the general infinite system (33) cannot be completely understood with the use of the theory of linear operators of Fredholm or Noether type.

*REFERENCES*


**1.** *Fedorov F. M.* Periodic Infinite Systems of Linear Algebraic Equations [in Russian]. Novosibirsk: Nauka, 2009.
**2.** *Fedorov F. M.* Periodic Infinite Systems of Linear Algebraic Equations and Their Applications [in Russian]. Novosibirsk: Nauka, 2011.
**3.** *Kantorovich L. V. and Akilov G. P.* Functional Analysis. Oxford and New York: Pergamon Press, 1982.
**4.** *Krasnov M. L.* Integral Equations. Introduction to the Theory [in Russian]. Moscow: Nauka, 1975.
**5.** *Kreĭn S. G.* Linear Equations in Banach Spaces [Russian]. Moscow: Nauka, 1971.
**6.** *Nöther F.* Über eine Klasse singulärer Integralgleichungen // Math. Ann. 1921. V. 82. P. 42–63.
**7.** *Fedorov F. M. and Osipova T. L.* A boundary method in problems with variable boundary conditions // Mat. Zametki YaGU. 2005. V. 12, N 1. P. 116–120.
**8.** *Fedorov F. M.* The Boundary Method for Solving Applied Problems of Mathematical Physics [in Russian]. Novosibirsk: Nauka, 2000.
**9.** *Cooke R.* Infinite Matrices and Sequence Spaces [Russian translation]. Moscow: Fizmatgiz, 1960.
**10.** *Godunov S. K. and Ryaben′kiĭ V. S.* Theory of Difference Schemes. An Introduction. Amsterdam etc.: North-Holland, 1964.
**11.** *Fedorov F. M., Ivanova O. F., and Pavlov N. N.* Convergence of reduction method and compatibility between infinite systems // Vestnik of NEFU. 2014. V. 11, N 2. P. 14–21.
**12.** *Gokhberg I. Ts. and Feldman I. A.* The Equations in the Verification and Projection Methods for Their Solution [in Russian]. Moscow: Nauka, 1971.
**13.** *Fedorov F. M.* On the theory of periodic infinite system of linear algebraic equations // Mat. Zametki YaGU. 2013. V. 20, N 1. P. 141–152.
**14.** *Fedorov F. M.* A new approach of studying real roots of a polynomial equation // Mat. Zametki YaGU. 2003. V. 10, N 1. P. 105–113.
**15.** *Fedorov F. M. and Abramova M. E.* Of solving algebraic equations of infinite power of a generalized method of Bernoulli // Mat. Zametki YaGU. 2004. V. 11, N 2. P. 80–88.
**16.** *Holubovski V. M.* Algebraic Properties of Groups of Infinite Matrices [in Russian], Dis. Dokt. Fiz.-Mat. Nauk, LOMI, St. Petersburg, 2007.


*September 20, 2015*

F. M. Fedorov; O. F. Ivanova; N. N. Pavlov
Ammosov North-Eastern Federal University, Yakutsk, Russia
`foma_46@mail.ru`

*Yakutian Mathematical Journal*
*October—December, 2015. Vol. 22, No. 4*

*UDC 517.9*

# FORMAL CLASSIFICATION OF GENERIC GERMS OF SEMI–HYPERBOLIC MAPPINGS

**P. A. Shaikhullina**

**Abstract.** We examine germs of a semi-hyperbolic mapping, i.e., two-dimensional holomorphic mappings one of whose multipliers is parabolic and the other is hyperbolic. Some formal classification of generic semi-hyperbolic germs is obtained and a theorem on semiformal normalization is proven.

**Keywords:** semi-hyperbolic germ, formal classification, semiformal normalization

## Introduction

The problem of an analytic classification of germs of vector fields (mappings) was stated by Poincaré [1–4]. Its version is the so-called normalization problem, i.e., the problem of reducing a germ to the simplest form (the normal form) by an analytic change of coordinates. Both these problems were solved by Poincaré [1–5], Siegel [6], and Bryuno [7] mainly. Only Siegel-type germs with resonances or pathologically close to them remain unstudied (see [8]). The essential achievements were made in the 1980s for these "particular" cases by Yoccoz [9] for the "Liouville" germs and by Voronin [10], Ecalle [11], Martinet and Ramis [12, 13] for the resonance germs. It turns out that an obstacles to a normalization of a germ are the so-called "functional invariants." And even more, they contain complete information about the "analytic" type of a germ.

Functional invariants of analytic classification were initially constructed for germs of parabolic mappings (one-dimensional holomorphisms tangent to the identity) [10, 11, 14], and for the resonances of saddle [14] and saddle-node [15] germs of vector fields. After that functional invariants were found in many other problems of analytic classification [16–19].

One of the methods for constructing functional invariants is as follows: A punctured neighborhood about a fixed (singular) point is covered by sectorial domains. At each of them we can construct an analytic change of variables normalizing a germ.

The transition functions of an obtained "normalizing" atlas are the required functional invariant. Hence, the problem on the sectorial normalization, i.e., the problem on normalization of a germ on a domain for which a fixed (singular) point is not interior but boundary is the first and most important stage of solving the problem of analytic classification of resonance germs. As is noted above, such a normalization was realized for one-dimensional resonance mappings and saddle and saddle-node resonance vector fields. The next objects in complexity are two-dimensional mappings and three-dimensional vector fields. The germs of two-dimensional mappings and three-dimensional vector fields (under very strong constraints separating subsets of the codimension infinity in the space of germs) are examined in [20] and the above program for these germs was completely realized there. However, generic resonance cases are not studied well. At present, the results on normalization probably are obtained only for the so-called semi-hyperbolic germs, i.e., the germs of two-dimensional mappings one of whose multipliers (with the modulus not equal to zero or 1) is hyperbolic, and the second is parabolic (it is equal to 1). Namely, the theorem on sectorial normalization of germs for a formal equivalence class of a particular semi-hyperbolic germ is proven in [15]. In what follows, the transition functions of a normalizing atlas are assumed to use in functional invariants of an analytic classifications of germs of the above class. We should observe that the class treated in [15] consists of germs whose fluxes (of sufficiently high order) are normalized at all points of their invariant hyperbolic submanifold. The presence of this "preliminary" normalization is essential for constructing sectorial normalizing mappings. However, the possibility of such preliminary normalization is not a direct consequence of the formal reducibility of a germ to its normal form.

Some particular results for germs of semi-hyperbolic mappings about the existence of a holomorphic ("sectorial") central manifold are obtained in [21, 22].

At the present article we consider generic semi-hyperbolic germs. We plan to obtain their analytic classification in accord with the above arguments. The first step needed for constructing a sectorial normalization is to establish their formal classification and "preliminary" normalizability.

## 1. The Classes of the Germs $\mathscr{SH}$ and $\mathscr{SH}_0$. A Theorem on Formal Classification

Definition 1. The germ of a holomorphism $F : (\mathbb{C}^2, 0) \to (\mathbb{C}^2, 0)$ is referred to as *semi-hyperbolic* whenever one of its multiplier is equal to 1 and the latter is hyperbolic, i.e.,

$$F(x, y) = (x + \cdots, \Lambda y + \cdots), \quad \text{where } |\Lambda| \neq 0, 1.$$

A semi-hyperbolic germ is called *generic* if the constant $c$ in the decomposition

$$F(x, y) = (x + cx^2 + \cdots, \Lambda y + \cdots)$$

is not zero.

The class of generic semi-hyperbolic diffeomorphisms is denoted by $\mathscr{SH}$. Note that $F \in \mathscr{SH}$ has an invariant analytic submanifold corresponding to a hyperbolic multiplier (Hadamard–Perron theorem [23]). It can be straightened by a holomorphic change of variables. Hence, without loss of generality, we can assume that this invariant submanifold is the straight line $\{x = 0\}$. The class of germs $F \in \mathscr{SH}$ whose invariant submanifold agrees with the straight line $\{x = 0\}$ is denoted by $\mathscr{SH}_0$.

Note that a germ of $\mathscr{SH}_0$ is resonance. By the Poincaré–Dulac theorem [8] a formal change of variables reduces this germ to the form

$$(x, y) \mapsto (f(x), yK(x)).$$

However, further normalization for these germs is also possible. Namely, the first component $x \mapsto f(x)$ is a parabolic mapping. The formal classification of parabolic mappings is well known [10]: in the typical case a formal equivalence class of such germ is defined only by a numerical modulus. It is convenient to employ the unit time shift $g^1_{v_a}$ along the vector field

$$v_a(x) = \frac{x^2}{1 + ax} \frac{\partial}{\partial x}, \quad a \in \mathbb{C}.$$

Next, there is an affinity between the results on classification of mappings and vector fields. A formal classification of vector fields with a singular point of saddle-node type in the two-dimensional case is exposed in [17]. For generic vector fields with such a point (the unit time shift along this vector field is a semi-hyperbolic mapping) the formal classification has three numerical parameter. So for generic semi-hyperbolic mappings we can take the unit time shifts along the formal normal forms (FNFs) of saddle-node vector fields from [17] as FNFs. However, it is more convenient for us to employ other FNFs which as before depend on three parameters. Moreover, in order to provide a formal classification of germs of semi-hyperbolic mappings from a classification of saddle-node vector fields, we need to prove a theorem on a formal inclusion of germs of semi-hyperbolic mappings into the flux which is a nontrivial task.

Moreover, as before we can note that for further studies of semi-hyperbolic mappings we need even more than just reducibility to formal normal form. However, both Poincaré-Dulac and the theorem on a formal inclusion in the flux do not give necessary results.

Hence, we prove the theorem on formal classification by direct arguments and obtain the theorem on formal inclusion in the flux (even in a stronger "semiformal" version) as a simple consequence. In this case we establish essentially stronger result that a germ of class $\mathscr{SH}_0$ is reducible to the FNF by a semiformal change.

**Theorem 1.** *Let $F \in \mathscr{SH}_0$. Then there exist complex numbers $a$, $\lambda$, and $\beta$ such that $F$ is formally equivalent to the unit time shift $F_{a,\lambda,\beta} = g^1_{\omega_{a,\lambda,\beta}}$ along the vector field*

$$\omega_{a,\lambda,\beta} = v_a(x) + y(\lambda + \beta v_a'(x))\frac{\partial}{\partial y}, \quad (a, \lambda, \beta) \in \mathbb{C}^3, \tag{1}$$

*where $v_a(x) = \frac{x^2}{1+ax}\frac{\partial}{\partial x}$, $\operatorname{Im}\lambda \in [0; 2\pi)$, and $\operatorname{Re}\lambda \neq 0$. Moreover, the normalizing transform can be chosen to be formal in the variables $x$ and analytic in $y$. (We call transformations of this type semiformal).*

Remark 1. We demonstrate below that the FNF is unique, i.e., for a given germ $F \in \mathscr{SH}_0$, the parameters $a$, $\lambda$, and $\beta$ of FNF are defined uniquely and the normalizing change is unique to within a "superposition" multiplication by the shift $g^t_{\omega_{a,\lambda,\beta}}$ for a fixed time $t \in \mathbb{C}$ and the delation $(x, y) \mapsto (x, ky)$, $k \neq 0$.

**Corollary 1.** *Let $F_{a,\lambda,\beta}$ be a formal normal form of a germ $F \in \mathscr{SH}_0$. Then, for every $N \in \mathbb{N}$, there exists a germ $\mathscr{G} \in \mathscr{SH}_0$ analytically equivalent to $F$ and such that*

$$\mathscr{G}(x, y) - F_{a,\lambda,\beta}(x, y) = O(x^N) \quad \text{as } x \to 0.$$

**Corollary 2.** *For a given germ $F \in \mathscr{SH}_0$, we can construct a formal vector field $v$ with a singular point of saddle-node type such that $F = g^1_v$.*

We use two stages to prove the theorem on semiformal classification. In Section 2 (the former stage) we reduce a semi-hyperbolic germ to the so-called preliminary formal normal form (PFNF) by the method of successive approximations with the use of semiformal change of variables. On the latter stage in Section 3, a PFNF is reduced to a polynomial form, and the end of the proof is in Section 4. In Section 5 we show uniqueness of the formal normal form (and "almost uniqueness" of a normalizing formal change). Corollaries 1 and 2 are proven in Sections 6 and 7, respectively.

## § 2. Preliminary Formal Normal Form

Assume that $\mathscr{F}$ is a germ of class $\mathscr{SH}_0$ and $F$ is its representative. First, we transform $F$ to a simpler form. Actually, it is a classical Poincaré–Dulac formal normal form. However, we justify reducibility to this form by "semiformal" changes of coordinates.

**Lemma 1.** *There exists a formal change of variables reducing a generic semi-hyperbolic mapping $F$ from $\mathscr{SH}_0$ to a form called preliminary formal normal form (PFNF)*

$$F_0(x,y) = (f(x), yK(x)), \quad \text{where } f(x) = x + x^2 + O(x^3),\ K(x) = \Lambda + O(x). \quad (2)$$

*A normalizing change of variables in this case can be chosen in the form*

$$(x,y) \mapsto \Big(\sum_i \alpha_i(y)x^i; \sum_j \beta_j(y)x^j\Big), \quad i,j = 0,1,\dots,$$

*with $\alpha_i, \beta_j$ analytic functions defined for all $i, j$ in some fixed neighborhood about the origin.*

Proof. 0. For a germ $\mathscr{F}$ of the class $\mathscr{SH}_0$, the line $\{x = 0\}$ is invariant and thereby $F(\{x = 0\}) = \{x = 0\}$. Represent the mapping $F$ by its Hartogs series in the polydisk $S_{rR} = \{(x,y) : |x| < r,\ |y| < R\}$ as follows:

$$F(x,y) = (xa_1(y) + x^2a_2(y) + \cdots, b_0(y) + b_1(y)x + \cdots),$$

where the functions $a_i(y)$, $b_j(y)$, $i = 1,2,\dots$, $j = 0,1,\dots$, are holomorphic. They satisfy the conditions $a_1(0) = 1$, $a_2(0) \neq 0$, $b_0(0) = 0$, $b_0'(0) = \Lambda$, and $|\Lambda| \neq 0, 1$. Below we use similar Hartogs series expansions and follow only the parameter $R$ (the radius of the disk of convergence in which all coefficients of the series are analytic).

The restriction $f$ of $F$ to the line $\{x = 0\}$ is a hyperbolic mapping $y \mapsto b_0(y)$. By the Schröder theorem [24], it can be linearized by an analytic change of variables. Thus, without loss of generality we can assume that $b_0(y) = \Lambda y$.

1. Normalization of the coefficient $a_1$. The analytic (on $S_{rR}$) change of variables $H_1(x,y) = (k(y)x, y)$ reduces $F$ to the form

$$F_1(x,y) = H_1^{-1} \circ F \circ H_1(x,y) = \left(xa_1(y)k(y)\frac{1}{k(\Lambda y + \cdots)} + \cdots, \Lambda y + \cdots\right).$$

We want to choose $k(y)$ such that

$$a_1(y)\frac{k(y)}{k(\Lambda y)} = 1. \quad (3)$$

If $|\Lambda| > 1$ then the holomorphic solution to (3) is representable as

$$k(y) = \prod_{n=1}^{\infty} a_1\left(\frac{y}{\Lambda^n}\right).$$

In the case of $0 < |\Lambda| < 1$, the holomorphic solution to (3) is the function

$$k(y) = \prod_{n=0}^{\infty} \frac{1}{a_1(\Lambda^n(y))}.$$

The convergence of infinite products and their holomorphy in the disk $\{|y| < R\}$ follows from the corresponding theorems of complex analysis. Thus, using some holomorphic (on $S_{rR}$) change of variables we can assume that the coefficient $a_1$ is equal to 1. In what follows we suppose that $a_1 \equiv 1$.

Remark 2. Note that after the above change of variables the second parameter of the polydisk $S_{rR}$ (the radius of convergence of the coefficients of the expansion of a normalizing mapping into the Hartogs series) remains the same.

2. Normalization of $a_2$ and $b_1$. Consider the transform

$$H_2(x,y) = (x + \alpha(y)x^2, y + \beta(y)x)$$

and its inverse

$$H_2^{-1}(x,y) = (x - \alpha(y)x^2 + O(x^3), y - \beta(y)x + O(x^2)).$$

In this case, as $x \to 0$, we have

$$\begin{aligned} F_2(x,y) &= H_2^{-1} \circ F_1 \circ H_2(x,y) \\ &= (x + x^2(\alpha(y) - \alpha(\Lambda y) + a_2(y)) + O(x^3), \\ &\Lambda y + x(\Lambda\beta(y) - \beta(\Lambda y) + b_1(y)) + O(x^2)). \end{aligned}$$

We want to solve the two equations

$$\alpha(\Lambda y) - \alpha(y) = a_2(y), \tag{4}$$

$$\beta(\Lambda y) - \Lambda\beta(y) = b_1(y). \tag{5}$$

This allow us to simplify the arguments essentially, more exactly, to equate the corresponding coefficients of the series $F_2$ to zero. Consider (4). Represent $a_2(y)$ and $\alpha(y)$ as a power series, i.e.,

$$\alpha(y) = \sum_{k=0}^{\infty} \alpha_k y^k, \quad a_2(y) = \sum_{k=0}^{\infty} a_{2,k} y^k.$$

The former series here is the expansion of $\alpha$ and the second converges in the disk $\{|y| < R\}$. Inserting these expressions into the equation and comparing the right-hand and left-hand sides, we infer

$$(\Lambda^k - 1)\alpha_k = a_{2,k}, \quad k = 0, 1, 2, \dots. \tag{6}$$

For every $k \neq 0$, the equation (6) is solvable and a solution for $k = 0$ exists if and only if $a_{2,0} = 0$. Repeating the same arguments for (5), we obtain the following expansions of $\beta_k$ and $b_{1,k}$:

$$\beta(y) = \sum_{k=0}^{\infty} \beta_k(y), \quad b_1(y) = \sum_{k=0}^{\infty} b_{1,k}(y)$$

and the system

$$(\Lambda^k - \Lambda)\beta_k = b_{1,k}, \quad k = 0, 1, 2, \dots,$$

whose equations are solvable for all $k \neq 1$. For $k = 1$, a necessary condition of solvability is that $b_{1,1}$ is equal to zero. So, making the change of variable of the above form we can assume that $a_2$ is a constant and $b_1$ has a linear form. Note that the genericity conditions imply that $a_2(0) \neq 0$. Hence, without loss of generality, we can assume that $a_2(y) = \text{const} \neq 0$ and $b_1(y) = b_{1,1}y$. At last, using the additional transformation $(x, y) \mapsto (mx, y)$, we can obtain that $a_2(y) = 1$.

Since $a_2(y)$ and $b_1(y)$ are analytic in the disk $\{|y| < R\}$, the series for $\alpha(y)$ and $\beta(y)$ converge in this disk. Hence, the transformation $H$ constructed is holomorphic on $S_{rR}$ and the remark of the previous section remains valid.

3. Induction. Let the mapping $F_{N-1}$ on some step be of the form

$$F_{N-1}(x,y) = (x + x^2 + \alpha_3 x^3 + \cdots + \alpha_{N-1}x^{N-1} + \alpha_N(y)x^N + O(x^{N+1});$$
$$y(\Lambda + \beta_1 x + \cdots + \beta_{N-2}x^{N-2}) + \beta_{N-1}(y)x^{N-1} + O(x^N)),$$

where $\alpha_3, \ldots, \alpha_{N-1}, \beta_1, \ldots, \beta_{N-2}$ are constants and the functions $\alpha_N(y)$ and $\beta_{N-1}(y)$ are holomorphic on $\{|y| < R\}$. We will use the transform

$$H_N(x,y) = (x + \alpha(y)x^N, y + \beta(y)x^{N-1}).$$

As on the previous step, we can make the coefficient $\alpha_N$ to be equal constant and $\beta_{N-1}$ linear in the variable $y$; the mapping $H_N(x,y)$ is holomorphic on $S_{rR}$ and Remark 2 remains valid.

4. Conclusion. Making the change of variables from the previous section for every $N$, we can construct a sequence of changes of variables $H_N$; each of them "improves" a pair of "monomials" from the expansion of a normalizable transform in the Hartogs series. Note that the stabilization of "monomials" in the expansion of the superposition $\mathscr{H}_N = H_1 \circ H_2 \circ \cdots \circ H_N$ holds. The same holds with the normalizable mapping $F_N = \mathscr{H}_N^{-1} \circ F \circ \mathscr{H}_N$. Hence, the infinite superposition $\mathscr{H} = \lim_{N\to\infty} \mathscr{H}_N$ converges in the space of formal series in the variable $x$ with coefficients holomorphic in $y$ in the disk $\{|y| < R\}$ and is an unknown normalizing semiformal change of variables, since the series $F_N$ converge to PFNF (2) in the same space of series. $\square$

Remark 3. The peculiarities of the construction imply that an infinite superposition reducing $F$ to PNF is a formal series in the variable $x$ with coefficients holomorphic in the same neighborhood about the origin .

## § 3. The Second Step of Formal Normalization

Above, we have reduced the generic semi-hyperbolic mapping to (2). The secondary normalization reducing it to FNF is realized with the help of the change

$$H(x,y) = (p(x), yq(x)), \tag{7}$$

$$p'(0) = 1. \tag{8}$$

This change reduces PFNF (2) to the form $\widetilde{F} = H^{-1} \circ F_0 \circ H = (\tilde{f}(x), \widetilde{K}(x)y)$, where

$$\tilde{f}(x) = p^{-1} \circ f \circ p, \quad \widetilde{K}(x) = K \circ p \cdot \frac{q}{q \circ \tilde{f}}. \tag{9}$$

Remark 4. The values

$$f'''(0) = \tilde{f}'''(0), \quad K(0) = \widetilde{K}(0), \quad K'(0) = \widetilde{K}'(0) \tag{10}$$

do not change after the change of coordinates (7), (8) (which are invariants of formal classification of generic semi-hyperbolic mappings). The latter is obvious for the second and third equalities and the first is demonstrated in [10] (but can be easily checked by the direct calculations).

Recall that $f$ is a generic parabolic mapping and $\tilde{f}$ is the adjoint of $f$. The formal classification of parabolic mappings implies ensures (see [10]) there exists a constant $a$ and a formal coordinate change of $p(x)$ reducing $f$ to $\tilde{f} = g^1_{v_a}(x)$, where $g^1_{v_a}(x)$ is a unit time shift along the vector field (1). Moreover, for $f(x)$ of the form $f(x) = x + x^2 + \cdots$ (the first component of the mapping (2) satisfies this condition), the change $p(x)$ satisfies (8). Next, assign $\varphi = \log q$ and $k = \log\big(\frac{\widetilde{K}}{K\circ p}\big)$. In this case the second equation in (9) takes the form

$$-\varphi \circ \tilde{f} + \varphi = k. \tag{11}$$

**Lemma 2.** *For every formal parabolic mapping $\tilde{f}$ and a function $k(x)$ such that $k(0) = k'(0) = 0$, there exists a unique formal solution $\varphi(x)$ to (11) such that $\varphi(0) = 0$.*

Proof. Let

$$\tilde{f} = x + cx^2 + \cdots, \; c \neq 0, \quad k(x) = \sum_{j=2}^{\infty} k_j x^j; \quad \varphi(x) = \sum_{j=1}^{\infty} \varphi_j x^j.$$

Inserting these expansions in (11), we infer

$$\sum_{j=1}^{\infty} \varphi_j x^j - \sum_{j=1}^{\infty} \varphi_j (x + cx^2 + \cdots)^j = \sum_{j=2}^{\infty} k_j x^j.$$

Equating the coefficients of $x^j$, $j = 1, 2, \dots,$ we arrive at the infinite system of equations

$$-c(j-1)\varphi_{j-1} + \cdots = k_j, \quad j = 2, 3, \dots,$$

where the dots stand for summands depending on $\varphi_s$ with the numbers $s$ less than $j-1$. Since $c \neq 0$ for a generic parabolic mapping, these equations are solvable successively. It gives the existence and uniqueness as well of a formal solution. □

**Lemma 3.** *For all parabolic diffeomorphisms $f(x) = x + x^2 + \cdots$ and $p(x) = x + \cdots$ and all $K$ and $\widetilde{K}$ satisfying (10), there exists a formal solution $q$ to the second equation of the system (9) such that $q(0) = 1$. This solution is unique (to within a multiplication by a constant ). In particular, every mapping $F \in \mathscr{S}\mathscr{H}_0$ of the form (2) is reducible to the form*

$$\widetilde{F}_{a,\lambda,b}(x, y) = \big(g^1_{v_a}(x), y(\Lambda + bx)\big), \quad \Lambda = e^{\lambda}$$

*by a formal change of the form (7).*

Proof. The former claim results from Lemma 2 and the latter from the former. □

## § 4. The End of the Proof of the Theorem on Formal Classification

Let $F_{a,\lambda,\beta} = g^1_{\omega_{a,\lambda,\beta}}(x, y)$ be the unit time shift along the vector field $\omega_{a,\lambda,\beta}$. We can solve the system of equations

$$\begin{cases} \dot{x} = v_a(x), \\ \dot{y} = y(\lambda + \beta v'_a(x)). \end{cases}$$

Let $(x(t), y(t))$ be its solution with the initial data ($x(0) = x_0$ and $y(0) = y_0$). We have

$$x(t) = g^t_{v_a}(x_0), \quad \frac{dy}{dx} = \lambda y \frac{1}{v_a(x)} + y\beta \frac{v'_a(x)}{v_a(x)}.$$

In this case,

$$dy = \lambda y\, dt + d(y \log v_a(x)),$$

and

$$y(t) = y_0 e^{\lambda t} \left( \frac{v_a(x_0)}{v_a(x(t))} \right)^{\beta}.$$

Assigning $f_a = g^1_{v_a}$, we infer

$$F_{a,\lambda,\beta}(x_0, y_0) = (f_a(x_0), e^{\lambda} y_0 (f'_a(x_0))^{\beta}).$$

At last, if

$$\Lambda = e^{\lambda}, \quad \Lambda(x) = \Lambda(f'_a(x))^{\beta},$$

then $F_{a,\lambda,\beta} = (f_a(x), y\Lambda(x))$.

Calculate the invariants (10) for $F_{a,\lambda,\beta}$ which are the parameters $a$, $\Lambda = e^{\lambda}$, $\beta$. By Lemma 3, $F_{a,\lambda,\beta}$ is reduced to $\widetilde{F}_{a,\lambda,\beta}$ by a change of the form (7). In accord with the constructions of this and previous sections, the initial mapping $F$ by semiformal changes is reduced to some form $\widetilde{F}_{a,\lambda,\beta}$. Hence, the initial mapping $F$ can be reduced to the form $F_{a,\lambda,\beta}$ by semiformal changes of variables. The theorem on formal classification is proven.

## § 5. Uniqueness of FNF and Normalizing Formal Coordinate Change

Assume that $F = F_{a,\lambda,\beta}$ and $F' = F_{a',\lambda',\beta'}$ are two formal normal forms from the theorem on formal classification and the formal coordinate change $H$ conjugates $F$ and $F'$, i.e.,

$$H \circ F = F' \circ H. \tag{12}$$

Note that the lines $\{x = 0\}$ and $\{y = 0\}$ are invariants for the formal normal forms (the former line is their common invariant hyperbolic submanifold and the latter the central manifolds). Hence, the coordinate change $H$ must preserve them. In this case the change $H$ is of the form

$$H(x, y) = (xa(x, y); yb(x, y)), \tag{13}$$

where $a(x, y)$ and $b(x, y)$ are formal power series and $a(0, 0) \cdot b(0, 0) \neq 0$. Therefore, the formal mapping $h(x) = xa(x, 0)$ conjugates the mappings $f_a(x)$ and $f_{a'}(x)$, i.e.,

$$h \circ f_a = f_{a'} \circ h. \tag{14}$$

Due to uniqueness of FNF of parabolic mappings [10], the parameters $a$ and $a'$ coincide. Similarly, the formal diffeomorphism $g(y) = yb(0, y)$ conjugates the restrictions of $F$ and $F'$ to the line $\{x = 0\}$ and thus

$$g(\Lambda y) = \Lambda' g(y). \tag{15}$$

Differentiating (15) with respect to $y$ and taking $y = 0$, we obtain $\Lambda = \Lambda'$ and the condition $\operatorname{Im} \lambda, \operatorname{Im} \lambda' \in [0; 2\pi)$ validates the inequality $\lambda = \lambda'$. Comparing the coefficients of $x^2$ in (14) in accord with the equality $a = a'$, we have

$$h'(0) = a(0, 0) = 1. \tag{16}$$

Inserting (13) in (12) and calculating the coefficient of $xy$ in the second component of the equality obtained, we derive from (16) that $\beta = \beta'$. The proof of uniqueness of FNF is completed. To study the question of uniqueness of a normalizing change, is suffices to examine formal coordinate changes preserving FNF. Thus, we need to find all formal changes $H(x, y)$ satisfying (12) (in the case of $F = F'$). Repeating the above arguments and using the same notations, we establish the same equalities (14) and (15), where $a = a'$, $\Lambda = \Lambda'$. The equality (14) implies that $h$ commutes with $f_a$. As is demonstrated in [10], all mappings $h$ formally commuting with $f_a$ have the form $h = g^t_{v_a}$ for some $t \in \mathbb{C}$. Consider the mapping $H_t = g^t_{\omega_{a,\lambda,\beta}}$; since $F = g^1_{\omega_{a,\lambda,\beta}}$, $H_t$ and $F$ commute, i.e., $H_t \circ F = F \circ H_t$. In this case $H_t^{-1} \circ H$ commutes with $F$. So without loss of generality, we can assume that $h = \mathrm{id}$ (we can achieve it replacing $H$ with $H_t^{-1} \circ H$).

Next, (15) validates the linearity of $g$: $g(y) = ky$, $k \neq 0$. The linear mapping $L : (x, y) \mapsto (x, ky)$ also commutes with the FNF $F$. Repeating the above arguments, we can assume that the mapping $g$ coincides with the identity. At last, inserting (13) in (12) and using the condition "$H = \mathrm{id}$ on the straight lines $\{x = 0\}$ and $\{y = 0\}$," sequentially equating the coefficients of the powers $x^k$ (i.e., we actually repeat arguments of Section 2–4 and solving relevant (homogeneous in our case) equations), we obtain that $H \equiv \mathrm{id}$. Two corrections of the normalizing change made above ensure its uniqueness to within a dilation along the $y$-axis and a shift along the vector field $\omega_{a,\lambda,\beta}$.

## § 6. Proof of Corollary 1

Let $H(x, y) = (H_1(x, y), H_2(x, y))$ be a normalizing formal change constructed in the theorem on a formal classification. Note that every of the series $H_j(x, y)$, $j = 1, 2$, is representable as

$$H_j(x, y) = \sum_{k=0}^{\infty} c^j_k(y) x^k,$$

where the functions $c^j_k(y)$ are holomorphic in some fixed neighborhood of the origin. Consider the change

$$H_N(x, y) = (H_{1N}(x, y), H_{2N}(x, y)),$$

where

$$H_{jN}(x, y) = \sum_{k=0}^{N} c^j_k(y) x^k.$$

It is holomorphic and takes the germ $F$ in the germ of the required form.

## § 7. Proof of Corollary 2

Assume that $F$ is an arbitrary germ in $\mathscr{S}\mathscr{H}_0$, $F_{a,\lambda,\beta} = g^1_{\omega_{a,\lambda,\beta}}$ is its FNF, and $H$ is a semiformal change of coordinates conjugating $F$ and $F_{a,\lambda,\beta}$: $H \circ F = F_{a,\lambda,\beta} \circ H$. Assume also that $H^{-1}$ takes the vector field $\omega_{a,\lambda,\beta}$ in the vector field $\omega$: $H' \cdot \omega = \omega_{a,\lambda,\beta} \circ H$. In this case, $F = g^1_\omega$ and the proof is completed.

Remark 5. Since the converse of the semiformal change is also semiformal, the above vector field $\omega$ is "semiformal" and so its components are expandable in the Hartogs series (in the powers of $x$) with coefficients analytic in some fixed neighborhood about the origin.

## *REFERENCES*

**1.** *Poincaré H.* Memoire sur les courbes definies par une equation differentielle. I // J. Math. Pures Appl. 1881. V. 7. P. 375–422.
**2.** *Poincaré H.* Memoire sur les courbes definies par une equation differentielle. II // J. Math. Pures Appl. 1882. V. 8. P. 251–286.
**3.** *Poincaré H.* Memoire sur les courbes definies par une equation differentielle. III // J. Math. Pures Appl. 1885. V. 1. P. 167–244.
**4.** *Poincaré H.* Memoire sur les courbes definies par une equation differentielle. IV // J. Math. Pures Appl. 1886. V. 2. P. 151–217.
**5.** *Poincaré H.* Sur le problem des trois corps et les equations de la dinamique // Acta Math. 1890. V. XIII. P. 1–270.
**6.** *Siegel C. L.* Vorlesungen uber Himmelsmechanik. Berlin; Göttingen; Heidelberg: Springer-Verlag, 1957.
**7.** *Bryuno A. D.* Analytic form of differential equations // Tr. Mosk. Mat. Obshch. 1971. V. 25. P. 199–239.
**8.** *Arnol′d V. I.* Supplementary Chapters to the Theory of Ordinary Differential Equations [in Russian]. Moscow: Nauka, 1978.
**9.** *Yoccoz J. C.* Linearisation des germes de diffeomorphismes holomorphes de ($\mathbb{C}$, 0) // C. R. Acad. Sci. Paris. Ser. I Math. 1988. V. 306, N 1. P. 55–58.
**10.** *Voronin S. M.* Analytic classification of germs of conformal mappings $(\mathbb{C}, 0) \to (\mathbb{C}, 0)$ with identity linear part // Funktsional. Anal. i Prilozhen. 1981. V. 15, N 1. P. 1–17.
**11.** *Ecalle J.* Sur les fonctions resurgentes. Orsay: Publ. Math. d'Orsay, 1981.
**12.** *Martinet J. and Ramis J. P.* Problème de modules pour des équations différentielles non linéaires du premier ordre // Inst. Hautes Études Sci. Publ. Math. 1982. N 55. P. 63–164.
**13.** *Martinet J. and Ramis J. P.* Classification analytique des équations différentielles non linéaires resonnantes du premier ordre // Ann. Sci. École Norm. Supér. 1983. V. 16, N 4. P. 571–621.
**14.** *Malgrange B.* Travoux d'Ecalle et de Martinet-Ramis sur les systemes dinamique // Asterisque. 1982. N 582. P. 59–73.
**23.** *Voronin S. M. and Fomina P. A.* Sectorial normalization of semihyperbolic mappings // Vestn. Chelyab. Gos. Univ. Ser. Mat. Mekh. Inform. No. 16. 2013. N 28. P. 94–114.
**16.** *Voronin S. M. and Meshcheryakova Yu. I.* Analytic classification of generic degenerate elementary singular points of germs of holomorphic vector fields on the complex plane // Izv. Vyssh. Uchebn. Zaved. Mat. 2002. N 1. P. 13–16.
**17.** *Voronin S. M. and and Meshcheryakova Yu. I.* Analytic classification of saddle nodes // Tr. Mosk. Mat. Obshch. 2005. V. 66. P. 93–113.
**18.** *Voronin S. M.* Analytic classification of pairs of involutions and its applications // Funktsional. Anal. i Prilozhen. 1982. V. 16, N 2. P. 21–29.
**19.** *Voronin S. M., Ortiz-Bobadilla L., and Rosales-González E.* Thom's problem for orbital analytic classification of degenerate singular points of holomorphic vector fields in the plane // Dokl. Math. 2010. V. 82, N 2. P. 759–761.
**20.** *Ueda T.* Local structure of analytic transformations of two complex variables // Int. J. Math. Kyoto Univ. 1986. V. 26, N 2. P. 233–261.
**21.** *Ueda T.* Local structure of analytic transformations of two complex variables // Int. J. Math. Kyoto Univ. 1991. V. 31, N 3. P. 695–711.
**22.** *Voronin S. M.* Darboux–Whitney's problem and related questions // Adv. Sov. Math. Nonlinear Stokes phenomena. Providence, RI: Amer. Math. Soc., 1993. V. 14. P. 139–233.
**24.** *Arnol′d V. I. and Il′yashenko Yu. S.* Ordinary differential equations // Dynamical Systems-1 [in Russian]. Moscow: VINITI, 1985. P. pp. 7–150.
**25.** *Schröder E.* Uber iterierte Functionen // Ann. Math. 1871. V. 3. P. 296–322.

*October 10, 2015*

P. A. Shaikhullina
Chelyabinsk State University, Chelyabinsk, Russia
`fominapa@gmail.com`

*Yakutian Mathematical Journal*
*October—December, 2015. Vol. 22, No. 4*

*UDC 519.633*

# FINITE DIFFERENCE METHOD FOR THE INVERSE PROBLEM OF THE SIMULTANEOUS DETERMINATION OF THE RIGHT–HAND SIDE AND THE LOWEST COEFFICIENTS IN PARABOLIC EQUATIONS

**Ling-De Su**

**Abstract.** We propose a numerical scheme to solve the inverse problem of determining two lower coefficients that depends on time only in the parabolic equation. The time dependence of the right-hand side of a parabolic equation is determined using additional solution values at points of the computational domain. For solving the nonlinear inverse problem, linearized approximations in time are constructed using standard finite difference procedures in space. The results of numerical experiments are presented, confirming the capabilities of the proposed computational algorithms for solving the coefficients inverse problem.

**Keywords:** inverse problem, finite difference method, parabolic partial differential equation, identification of the coefficients

In this paper, we consider the inverse problem of finding the two coefficients $\gamma(t)$ and $f(t)$ in the following equation

$$u_t - u_{xx} + \gamma(t) \cdot u = f(t) \cdot g(x,t), \tag{1}$$

where $g(x,t)$ is a known analytic function, the solution $u(x,t)$ and the two coefficients $\gamma(t)$ and $f(t)$ are unknown.

Many inverse problems arise in engineering and mathematical sciences [1–4]. In a direct problem, it is required to find a solution that satisfies some given partial differential equation and some initial and boundary conditions. Different from direct problem, in inverse problems, the master equation, initial conditions and boundary conditions are not fully specified, instead, some additional information is available.

There are various kinds of inverse problems in physics: coefficient inverse problems (in which the equation is not specified completely as some equation coefficients are unknown), boundary inverse problems (in which boundary conditions are unknown), and evolutionary inverse problems (in which initial conditions are unknown) [5, 6]. In this paper, we focus on the inverse problem of the parabolic equations is an inverse coefficient problem.

Many inverse problems are formulated as non-classical problems for PDEs. In other words, most standard numerical methods cannot achieve good accuracy in solving this problems. In solving these problems approximately, emphasis is on the development of stable computational algorithms that take into account peculiarities of inverse problems [7, 8]. Several regularization methods have been developed for solving ill-conditioning problems [8–10].

Much attention is paid to the problem of determining the lower coefficient of a parabolic equation of second order. The existence and uniqueness of the solution for such an inverse problem and well-posed of this problem in various functional classes were examined [11, 12]. Numerical methods for solving the problem of the identification of the lower coefficient of parabolic equations are also considered in many works [13–15].

In this paper, we consider the problem of determining the two lower coefficient that depends only on time. Approximation in space is performed using standard finite difference. The main features of the nonlinear inverse problem are taken into account via a proper choice of the linearized approximation in time. Linear problems at a particular time level are solved on the basis of a special decomposition into new standard elliptic problems.

The layout of the article is as follows: In Section 1, we briefly introduce the formulation of the problem. In Section 2, we introduce the method and apply this method to the time-dependent problems. The results of numerical experiments are presented in Section 3. Section 4 is dedicated to a brief conclusion. Finally, some references are introduced at the end.

## 1. Formulation of the Problem

Let $\Omega$ be a bounded domain with a smooth boundary $\partial\Omega$. The direct problem is formulated as follows:

$$u_t - u_{xx} + \gamma(t) \cdot u = f(t) \cdot g(x,t), \quad x \in \Omega,\ 0 < t \leqslant T, \tag{2}$$

with initial condition

$$u(x,0) = u_0(x), \quad x \in \overline{\Omega}, \tag{3}$$

and the Dirichlet boundary condition

$$u(0,t) = \Phi_1(t), \quad u(l,t) = \Phi_2(t), \quad 0 < t \leqslant T, \tag{4}$$

where $\gamma(t)$, $f(t)$, $g(x,t)$ and $\Phi_i(t)$ $(i = 1,\ 2)$ are known functions, $u(x,t)$ is the solution of the second order parabolic equation (2), and the initial condition $u_0(x)$ is known.

In this paper, we consider the coefficient inverse problem, in (2), where the coefficients $\gamma(t)$ and $f(t)$ are unknown. The additional conditions are often formulated as:

$$\int\limits_{\Omega} u(x,t)\omega(x)\,dx = \varphi(t), \quad \int\limits_{\Omega} u(x,t)\chi(x)\,dx = \psi(t), \quad 0 < t \leqslant T, \tag{5}$$

where $\omega(x)$ and $\chi(x)$ are weight functions. Specifically, choosing $\omega(x) = \delta(x - x^*)$ and $\chi(x) = \delta(x - x^{**})$, $(x^*, x^{**} \in \Omega)$, where $\delta(x)$ is the Dirac $\delta$-function, from (5), we get

$$u(x^*,t) = \varphi(t), \quad u(x^{**},t) = \psi(t), \quad 0 < t \leqslant T. \tag{6}$$

The inverse problem of finding $u(x,t)$, $\gamma(t)$, and $f(t)$ from problems (2)–(4) and additional conditions (5) or (6) is well-posed. The corresponding conditions for existence and uniqueness of the solution are available in the above-mentioned articles.

In this paper, we consider only the numerical solution of these inverse problems with one dimension, omitting the theoretical issues of the convergence of an approximate solution to the exact one.

## 2. The Computational Algorithm

The inverse problem (2)–(5) (or (2)–(4), (6)) is nonlinear. The standard approach is based on the simplest approximations in time and involves the iterative solution of the corresponding nonlinear problem for the evaluation of the approximate solution at a new level. In this article, we apply such approximations in time that lead to linear problems for evaluating the solution at the new time level.

**2.1. The inverse problem.** We consider the inverse problem (2)–(4), (6) where $\Omega = [0, l]$. To solve the parabolic problem numerically, we introduce the grid in space

$$\overline{\omega}_h = \{x \mid x_i = ih,\ i = 0, 1, 2, \dots, M,\ Mh = l\},$$

for the time we also have

$$\overline{\omega}_\tau = \{t^n \mid t^n = n\tau,\ n = 0, 1, 2, \dots, N,\ N\tau = T\}.$$

For all but boundary grid nodes, we use the operator $D$ written as

$$Du = -\frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2}, \quad x \in \omega_h.$$

Finite difference approximations in space are employed. Using the fully implicit scheme for approximation in time and the operator $D$, the notation $u^n = u(x, t^n)$, $t^{n+1} = t^n + \tau$ $(n = 0, 1, 2, \dots, N-1)$, we obtain the following variational problem:

$$\frac{u^{n+1} - u^n}{\tau} + Du^{n+1} + \gamma^{n+1} \cdot u^n = f^{n+1} \cdot g^{n+1}, \quad x \in \omega_h, \tag{7}$$

the initial condition

$$u(x, 0) = u_0(x), \quad x \in \overline{\omega}_h, \tag{8}$$

and the boundary conditions

$$u(0,t) = \Phi_1(t), \quad u(l,t) = \Phi_2(t), \quad t \in \overline{\omega}_\tau, \tag{9}$$

the additional conditions (6) take the form:

$$u^n(x^*) = \varphi^n, \quad u^n(x^{**}) = \psi^n, \quad n = 0, 1, 2, \dots, N. \tag{10}$$

We use the following decomposition for the solution $u^{n+1}$ at the new time level

$$u^{n+1}(x) = y^{n+1}(x) + \gamma^{n+1} v^{n+1}(x) + f^{n+1} w^{n+1}(x). \tag{11}$$

To find $y^{n+1}(x)$, inserting (11) in (7), we employ the equation

$$\frac{y^{n+1} - u^n}{\tau} + Dy^{n+1} = 0, \quad n = 0, 1, \dots, N-1,\ x \in \omega_h, \tag{12}$$

with the boundary conditions $y(0,t) = \Phi_1(t)$, $y(l,t) = \Phi_2(t)$.

The functions $v^{n+1}(x)$ and $w^{n+1}(x)$ are determined from

$$\frac{v^{n+1}}{\tau} + Dv^{n+1} = -u^n, \quad n = 0, 1, 2, \dots, N-1,\ x \in \omega_h, \tag{13}$$

$$\frac{w^{n+1}}{\tau} + Dw^{n+1} = g^{n+1}, \quad n = 0, 1, 2, \dots, N-1,\ x \in \omega_h, \tag{14}$$

with the boundary condition $v(0,t) = v(l,t) = 0$ and $w(0,t) = w(l,t) = 0$.

To evaluate $\gamma^{n+1}$ and $f^{n+1}$, the addition conditions (10) are used; inserting (11) into (10), we get

$$\begin{aligned} \gamma^{n+1}v^{n+1}(x^*) + f^{n+1}w^{n+1}(x^*) &= \varphi^{n+1} - y^{n+1}(x^*), \\ \gamma^{n+1}v^{n+1}(x^{**}) + f^{n+1}w^{n+1}(x^{**}) &= \psi^{n+1} - y^{n+1}(x^{**}), \end{aligned} \tag{15}$$

where $x^*, x^{**} \in [0, l]$ and $x^* \neq x^{**}$. To solve $\gamma^{n+1}$ and $f^{n+1}$ from (15), we assume $v^{n+1}(x^*)w^{n+1}(x^{**}) - v^{n+1}(x^{**})w^{n+1}(x^*) \neq 0$, where $v^{n+1}$ and $w^{n+1}$ are determined from (13) and (14).

Thus, the computational algorithm for solving the inverse problem (2)–(4), (6) based on the linearized scheme (7)–(10) involves the solution of three standard grid elliptic equations for the auxiliary functions $y^{n+1}(x)$ from equation (12), $v^{n+1}(x)$ form equation (13) and $w^{n+1}(x)$ from equation (14), the further evaluation of $\gamma^{n+1}$ and $f^{n+1}$ from (15), and the final calculation $u^{n+1}(x)$ from the relation (11).

**2.2. The solutions of $\boldsymbol{y^{n+1}(x)}$, $\boldsymbol{v^{n+1}(x)}$, and $\boldsymbol{w^{n+1}(x)}$.** To calculate $u^{n+1}(x) = y^{n+1}(x) + \gamma^{n+1}v^{n+1}(x) + f^{n+1}w^{n+1}(x)$, $n = 0, 1, \ldots, N-1$, we should find $y^{n+1}(x)$, $v^{n+1}(x)$ and $w^{n+1}(x)$ from (12)–(14).

To find $y^{n+1}(x)$, $v^{n+1}(x)$ and $w^{n+1}(x)$, using the notation $y_i^n = y(x_i, t^n)$, $t^{n+1} = t^n + \tau$, $n = 0, 1, 2, \ldots, N-1$, $x_i = x_{i-1} + h$, $i = 1, 2, \ldots, M$, converting (12)–(14) to the following forms with the boundary conditions:

$$\begin{gathered} y_{i+1}^{n+1} - \left(2 + \frac{h^2}{\tau}\right) y_i^{n+1} + y_{i-1}^{n+1} + \frac{h^2}{\tau} u_i^n = 0, \quad i = 1, 2, \ldots, M-1, \\ y_0^{n+1} = \Phi_1(t^{n+1}), \quad y_M^{n+1} = \Phi_2(t^{n+1}), \end{gathered} \tag{16}$$

$$\begin{gathered} v_{i+1}^{n+1} - \left(2 + \frac{h^2}{\tau}\right) v_i^{n+1} + v_{i-1}^{n+1} + h^2 u_i^n = 0, \quad i = 1, 2, \ldots, M-1, \\ v_0^{n+1} = v_M^{n+1} = 0, \end{gathered} \tag{17}$$

$$\begin{gathered} w_{i+1}^{n+1} - \left(2 + \frac{h^2}{\tau}\right) w_i^{n+1} + w_{i-1}^{n+1} + h^2 g_i^{n+1} = 0, \quad i = 1, 2, \ldots, M-1, \\ w_0^{n+1} = w_M^{n+1} = 0. \end{gathered} \tag{18}$$

Using the following decompositions for the solutions $y_i^{n+1}$, $v_i^{n+1}$, and $w_i^{n+1}$:

$$\begin{gathered} y_i^{n+1} = \alpha_i^{n+1} \cdot y_{i+1}^{n+1} + \beta_i^{n+1}, \quad v_i^{n+1} = \alpha_i^{n+1} \cdot v_{i+1}^{n+1} + \gamma_i^{n+1}, \\ w_i^{n+1} = \alpha_i^{n+1} \cdot w_{i+1}^{n+1} + \delta_i^{n+1}, \ i = M-1, M-2, \ldots, 0, \end{gathered} \tag{19}$$

and $y_M^{n+1} = \Phi_2(t^{n+1})$, $v_M^{n+1} = w_M^{n+1} = 0$.

By combining (16)–(19), we can get:

$$\begin{gathered} \alpha_i^{n+1} = \frac{1}{2 + \kappa - \alpha_{i-1}^{n+1}}, \quad \beta_i^{n+1} = \frac{\kappa u_i^n + \beta_{i-1}^{n+1}}{2 + \kappa - \alpha_{i-1}^{n+1}}, \\ \gamma_i^{n+1} = \frac{h^2 u_i^n + \gamma_{i-1}^{n+1}}{2 + \kappa - \alpha_{i-1}^{n+1}}, \quad \delta_i^{n+1} = \frac{h^2 g_i^{n+1} + \delta_{i-1}^{n+1}}{2 + \kappa - \alpha_{i-1}^{n+1}}, \\ i = 1, 2, 3, \ldots, M. \end{gathered} \tag{20}$$

with the conditions $\alpha_0 = \gamma_0 = \delta_0 = 0$ and $\beta_0 = \Phi_1$, where $\kappa = \frac{h^2}{\tau}$. From (20) we can get all the $\alpha$, $\beta$, $\gamma$ and $\delta$, so we can get all the solutions of $y$, $v$, and $w$.

## 3. Numerical Examples

In this section we present numerical results to test the efficiency of the new scheme for solving the coefficient inverse problems. In the example, we put $x \in [0, 1]$ with the conditions

$$\begin{gathered} u_0 = 5\exp(-100(x-0.5)^2), \quad g(x) = 4(1-x)x, \\ u(0,t) = u(1,t) = 0, \quad 0 < t \leqslant T. \end{gathered} \tag{21}$$

The coefficients $\gamma(t)$ and $f(t)$ are taken in the forms

$$\gamma(t) = -\frac{t^3}{1+\exp\left(\zeta_1(t-0.7T)\right)}, \quad f(t) = \frac{0.1(T-t)}{1+\exp\left(\zeta_2((T-t)-0.8T)\right)}. \tag{22}$$

We consider the inverse problem with $h = \frac{1}{M}$, $\tau = \frac{T}{N}$, $M = 200$, $N = 400$. The observation points are $x^* = 0.5$ and $x^{**} = 0.1$. The solution of the direct problem at the observation point are depicted in Fig. 1(a), the solution at the final time moment $(T = 0.5)$ is presented in Fig. 1(b), with $\zeta_1 = 1000$, $\zeta_2 = 250$.

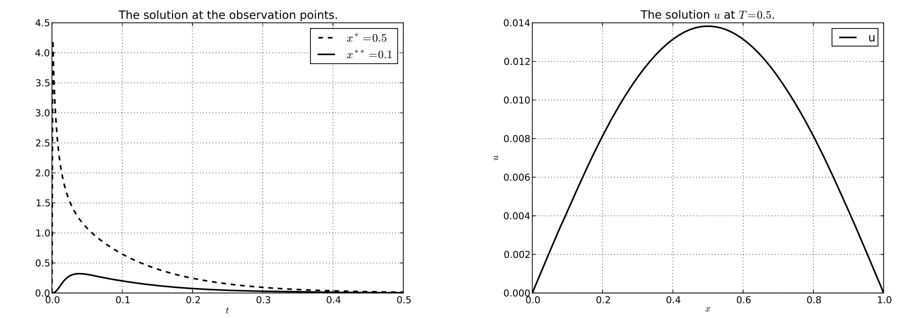



Fig. 1. The solution at the observation point and the solution at $T = 0.5$

We also give the solution of $\gamma(t)$ (Fig. 2(a)) and $f(t)$ (Fig. 2(b)) of the inverse problem with $\zeta_1 = 1000$, $\zeta_2 = 250$. The graphs show the very good accuracy and efficiency of the new approximate scheme.

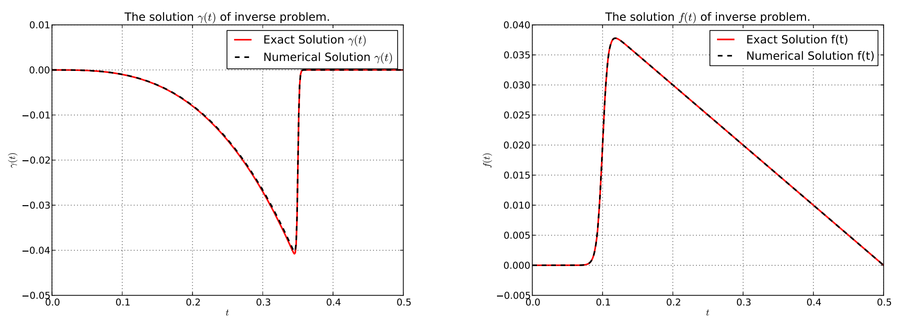



Fig. 2. The exact and numerical solutions of the inverse problem

The solution $\gamma(t)$ of the inverse problem with different $\zeta_1$ are present in Fig. 3(a). For large $\zeta_1$ (see Fig. 3(a)), $\gamma(t)$ approaches discontinuous functions with a discontinuity point at $t = 0.7T$.

The solution $f(t)$ of the inverse problem with different $\zeta_2$ are present in Fig. 3(b). For large $\zeta_2$ (see Fig. 3(b)), $f(t)$ also approaches discontinuous functions with a discontinuity point at $t = 0.2T$.

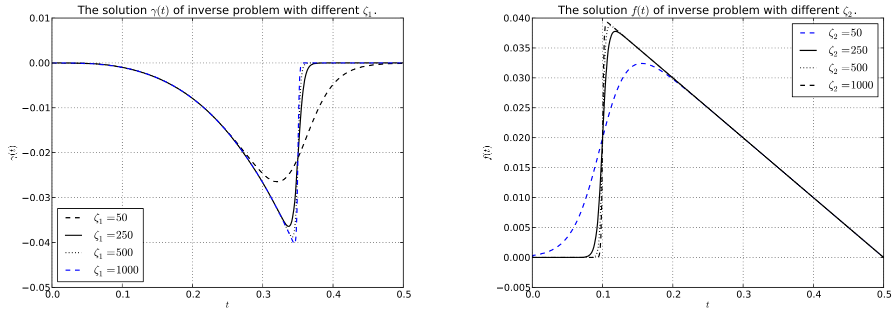



Fig. 3. The solutions of the inverse problem with different variables

## 4. Conclusions

In this paper, we proposed a numerical scheme for solving the inverse problems by the finite difference method. The numerical implementation using the linearized approximations in time and standard finite difference procedures in space, based on a decomposition of the approximate solution, where the transition to a new time level involves the solutions of three standard elliptic problems. Numerical solutions of the model problem demonstrate the convergence of the approximate solution of the inverse problem.

## 5. Acknowledgments

Professors V. I. Vasil′ev and P. N. Vabishchevich carefully reviewed this paper. As a result of their careful analysis, this paper was improved. The author expresses his thankfulness to them for helpful constructive comments.

## *REFERENCES*

1. *Colton D., Ewing R., and Rundell W.* Inverse Problems in Partial Differential Equations. Philadelphia, PA: SIAM, 1990.
2. *Engl H. W. and Rundell W.* Inverse Problems in Diffusion Processes. Philadelphia, PA: SIAM, 1995.
3. *Conca C., Manasevich R., Uhlmann G., and Vogelius M. S.* Partial differential equations and inverse problems // Contemp. Math. 2004. V. 362. P. 410.
4. *Muller T. G. and Timmer J.* Parameter identification in partial differential equations // Int. J. Bifurcation Chaos Appl. Sci. Eng. 2004. V. 14, N 6. P. 2053–2060.
5. *Samarskii A. A. and Vabishchevich P. N.* Numerical Methods for Solving Inverse Problems of Mathematical Physics. Utrecht: VSP, 2007.
6. *Jiang T. S., Li M., and Chen C. S.* The Method of particular solutions for solving inverse problems of a nonhomogeneous convection-diffusion equation with variable coefficients // Numerical Heat Transfer. Part A: Applications. , 2012. V. 61. P. 338–352.

**7.** *Hansen P. C.* The $L$-curve and its use in the numerical treatment of inverse problems // Computational Inverse Problems in Electrocardiology (ed. P. Johnston). (Adv. Comput. Bioeng. Ser.) Southampton: WITPress, 2000. P. 119–142.
**8.** *Hansen P. C.* Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems // Numer. Algorithms. 1994. V. 6, N 1. P. 1–35.
**9.** *Hansen P. C. and O'Leary D. P.* The use of the $L$-curve in the regularization of discrete ill-posed problems // SIAM J. Sci. Comput. 1993. V. 14, N 6. P. 1487–1503.
**10.** *Hao D. N.* A mollification method for ill-posed problem // Numer. Math. 1994. V. 68, N 4. P. 469–506.
**11.** *Isakov V.* Inverse Problems for Partial Differential Equations. Berlin: Springer-Verlag, 2006. (Appl. Math. Sci.; V. 127).
**12.** *Prilepko A. I., Orlovsky D. G., and Vasin I. A.* Methods for Solving Inverse Problems in Mathematical Physics. Berlin: Marcel Dekker, 2000.
**13.** *Dehghan M.* An inverse problem of finding a source parameter in a semilinear parabolic equation // Appl. Math. Modelling. 2001. V. 25, N 9. P. 743–754.
**14.** *Chan T. F. and Tai X. C.* Level set and total variation regularization for elliptic inverse problems with discontinuous coefficients // J. Comput. Phys. 2004. V. 193, N 1. P. 40–66.
**15.** *Vabishchevich P. N., Vasil′ev V. I., and Vasil′eva M. V.* Computational identification of the right-hand side of a parabolic equation // Comput. Math. Math. Phys. 2015. V. 55, N 9. P. 1015–1021.

*September 18, 2015*

Ling-De Su
Ammosov North-Eastern Federal University
Institute of Mathematics and Information Science, Yakutsk, Russia
sulingde@gmail.com