UZBEKISTAN ACADEMY OF SCIENCES
V.I.ROMANOVSKIY INSTITUTE OF MATHEMATICS

UZBEK MATHEMATICAL JOURNAL

Journal was founded in 1957. Until 1991 it was named by "Izv. Akad. Nauk UzSSR, Ser. Fiz.-Mat. Nauk". Since 1991 it is known as "Uzbek Mathematical Journal". It has 4 issues annually.

Volume 70 Issue 2 2026

Uzbek Mathematical Journal is abstracting and indexing by

SCOPUS **EBSCO** **CrossRef**
MathSciNet **Zentralblatt** **Math VINITI**

TASHKENT - 2026

Editorial Board

Editor in Chief

Sh.A. Ayupov – (Functional Analysis, Algebra), V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan), sh_ayupov@mail.ru

Deputy Editor in Chief

U.A.Rozikov – (Functional analysis, mathematical physics), V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan), rozikovu@yandex.ru

Managing editors

A.F.Aliyev, D.M.Akhmedov, – Managing editors of the Uzbek Mathematical Journal, V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan).

Honorary editors

Sh.A. Alimov – (Mathematical Analysis, Differential Equations, Mathematical Physics) National University of Uzbekistan (Uzbekistan)

A. Azamov – (Dynamical Systems, Game Theory, Differential Equations) V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)

M.M. Aripov – (Ordinary differential equations), National University of Uzbekistan (Uzbekistan)

A.A. Borubaev – Institute of Mathematics of the National Academy of Sciences of the Kyrgyz Republic. (Kyrgyztan)

V.I. Chilin – (Functional analysis), V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)

A. van Enter – (Probability and mathematical physics) University of Groningen (The Netherlands)

Sh.K. Formanov – (Probability Theory and Mathematical Statistics), V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)

R.N. Ganikhodjaev – (Functional analysis), National University of Uzbekistan (Uzbekistan)

N.N. Ganikhodjaev – (Functional analysis), V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)

F. Hiroshima – (Spectral and stochastic analysis, functional integration, Quantum field theory), Kyushu University (Japan)
D. Khadjiev – (Algebraic geometry, Associative rings and algebras, Differential geometry), National University of Uzbekistan (Uzbekistan)
M. Ladra – (category theory: homological algebra, nonassociative rings and algebras), University of Santiago de Compostella (Spain)
S.N. Lakaev – (Mathematical physics), Samarkand State University, Samarkand, (Uzbekistan)
A. Le Ny – (Probability and Statistical Mechanics) University Paris-Est (France)
Sh. Mirakhmedov – (Boundary Problems of Probability Theory, Discrete Problems of Probability Theory), V.I.Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
L. Ramero – (algebraic and arithmetic geometry, commutative rings and algebras), University of Lille (France)
F.A. Sukochev – (Functional analysis, Geometry), University of South Wales (Australia)
E. Zelmanov – (Agebra, Jordan Algebras, Infinite Discrete Groups, Profinite Groups), The Southern University of Science and Technology, Shenzhen, (China).

Editors

H. Akhadkulov – University Utara Malaysia, (Malaysia)
R.D. Aloev – National University of Uzbekistan, (Uzbekistan)
R.R. Ashurov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
A. Atamuratov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
R.B. Beshimov – National University of Uzbekistan (Uzbekistan)
R. Davronov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
A. Dzhalilov – Turin Polytechnic University in Tashkent, (Uzbekistan)
D.K. Durdiev – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
F.Kh. Eshmatov – New Uzbekistan University (Uzbekistan)
Z.K. Eshkuvatov – The Tashkent Institute of Irrigation and Agricultural Mechanization Engineers National Research University, (Uzbekistan)
F.H. Haydarov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
A.R. Hayotov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
G.I. Ibragimov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
R. Ibragimov – Imperial College Business School, (UK)
I.A. Ikromov – V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)

U.U. Jamilov	–	V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
B.E. Kanetov	–	Kyrgyz National University named after J.Balasagyn (Kyrgyzstan)
E. Karimov	–	Ghent University (Belgium)
O.N. Khakimov	–	V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
K. Khompysh	–	Al-Farabi Kazakh National University (Kazakhstan)
A.Kh. Khudoyberdiyev	–	V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
L. Kocinac	–	University of Nis (Serbia)
K.L. Kozlov	–	Moscow State University (Russia)
K.K. Kudaybergenov	–	Harbin Institute of Technology (China)
L. Liao	–	School of Mathematics and Statistics, Wuhan University, (China)
F.M. Mukhamedov	–	Department of Mathematical Sciences, College of Science, (UAE)
E.R. Oliveira	–	Instituto de Matemática e Estatística – UFRGS (Brazil)
B.A. Omirov	–	Harbin Institute of Technology, (China)
F. Ouaridi	–	Department of Geometry and Topology, University of Sevilla, Sevilla (Spain)
R. Parovik	–	Institute of Cosmophysical Research and Radio Wave Propagation (Russia)
A.V. Pskhu	–	Institute of Applied Mathematics and Automation, KBSC RAN, (Russia)
K.X. Rahimov	–	V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
A.A. Rakhimov	–	National University of Uzbekistan (Uzbekistan)
I.S. Rakhimov	–	Universite of Technology MARA, (Malaysia)
T.H. Rasulov	–	Bukhara State University, (Uzbekistan)
M. Ruzibaev	–	V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences (Uzbekistan)
K.M. Shadimetov	–	Tashkent State Transport University, (Uzbekistan)
O.Sh. Sharipov	–	National University of Uzbekistan, (Uzbekistan)
Y. Sheng	–	Mathematics School, Institute of Jilin University, (China)
B. Torebek	–	Institute of Mathematics and Mathematical Modeling (Kazakhstan)
D.A. Towers	–	School of Mathematical Sciences, Lancaster University (UK)
G. Urazboev	–	Urgench State University (Uzbekistan)
K. Van Bockstal	–	Ghent University (Belgium)
E.B. Yarovaya	–	Lomonosov Moscow State University, (Russia)
A.A. Zaitov	–	Tashkent University of Architecture and Civil Engineering (Uzbekistan)

Address: University str., 9, Tashkent 100174, Uzbekistan

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 5–14
DOI: 10.29229/uzmj.2026-2-1

On the accuracy of approximate solution scheme based on the Taylor formula for a Cauchy problem for ODEs

Abdullaev A., Azamov A., Ruziboev M., Tilavov A.

Abstract. In this work, we obtain explicit estimates for the accuracy of an approximation solution scheme for ODEs. The scheme is of the fourth order and based on the Taylor formula. Here we derive estimates with explicit constants, which are expressed by the derivatives of the right-hand side.

Keywords: Dynamical system, Cauchy problem, approximate solution, Taylor formula, Gronwall's lemma, solution accuracy, accuracy level, error estimation
MSC (2020): 26D05, 26E60

1. Introduction

In many applications of mathematics, we face the problem of approximately solving the Cauchy problem

$$\dot{x}(t) = f[x(t)], \quad x(0) = x_0 \tag{1.1}$$

Problem (1.1) is usually given for $x \in D \subset \mathbb{R}^d$, D where is a convex domain, $f : D \to \mathbb{R}^d$ is assumed to be smooth. In this work, we assume it is four times continuously differentiable. The effectiveness of an approximate solution scheme is defined by its accuracy. If $x_*(t)$ is the solution of problem (1.1) and $\tilde{x}(t)$ is the approximate solution in $(0 \le t \le T)$, then the level of accuracy is defined by

$$\sup_{0\le t\le T} |x_*(t) - \tilde{x}(t)| \le Ce^{LT}h^s \tag{1.2}$$

where L is the Lipschitz constant of f, h is the mesh size, s is the accuracy of the approximate solution scheme. There is extensive literature devoted to this problem. See for example [1, 2, 3, 4, 5]. The most common one step approximate solution schemes can be divided into two groups.

1. Approximate solution schemes based on the Taylor formula,
2. Runge-Kutta method schemes.

Generally, the Runge-Kutta method is preferred to the Taylor formula method [2]. This conclusion is explained by the necessity to calculate the derivatives of the function $f(x)$ in the approximate solution method based on the Taylor formula. In this article, we consider that approximate solution schemes based on the Taylor has a valuable property: it is possible to obtain accuracy of the method explicitly in terms of the derivatives of the right hand side of the system. Such schemes also have their positive aspects. Today, due to the development of computer algebra, it has become possible to automatically calculate the derivatives of a broad class of functions $f(x)$ [6, 7, 8].

There is another important reason for a deeper study of the Taylor method. Since the 1960s, numerical methods have been applied not only to approximate solutions of dynamic systems but also to study their properties, specifically in proving theorems that represent fixed properties [9, 10]. This requires a deductive proof of inequalities of the form (1.2), as is accepted in mathematics. From this perspective, approximate solution methods based on the Taylor formula are significantly more convenient than Runge-Kutta methods. [11] established such inequality for a scheme with accuracy level 3 in their article. In this article, for problem (1.1) with $s = 4$, the inequality (1.2) is proved, and a fixed expression is found for the corresponding coefficient C.

The issue of assessing the accuracy of approximate solutions is covered in the literature dedicated to numerical methods with varying degrees of detail. Here, we highlight several prominent monographs. In [12] the error assessment issue is not addressed. In the monographs, [5, 13, 14], and [4] only Runge-Kutta methods are examined. Part III of Bakhvalov's fundamental monograph [1] is devoted to approximate solution methods for ordinary differential equations. Specifically, in the first section

of this part, the Taylor method is discussed from a methodological point of view. The monograph [15] mentions the Taylor method in the second section of the second part, but the problem of estimating the error terms is not considered. It is only noted that if the first n terms of the Taylor formula are taken when constructing an approximate solution scheme, the error will be

$$\max_{0\le n\le N} |x_n - y(x_n)| = \mathrm{O}(h^n).$$

In this article, as noted above, an exact upper bound for the error is found and a strict proof of inequality (1.2) with explicit constants is provided. Such estimates are of vital importance when using numerical solutions to prove mathematical theorems. Notice that, in the last three decades computational methods being applied to obtain rigorous results in dynamical systems. Some of them have already solved long standing open problems, like the celebrated work of Tucker [16, 17], where numerical analysis combined with analytical to obtain the existence of Lorenz attractors. Many other methods have been devised to obtain various other properties, like the existence of homoclic trajectories; see, for example, [18, 19, 20] and references threin.

Here we set up a 4th order explicit numerical scheme, based on the Taylor formula and obtain explicit bounds for the its accuracy. Notice that our constant in (1.2) can be explicitly computed in terms of the right hand side of the equation.

2. Approximate solution scheme based on the Taylor formula

For a continuous function $f(x)$, a solution to problem (1.1) exists and is unique, but it is impossible to determine its domain of definition [21]. With this in mind, we begin the discussion with the following assumption.

Assumption 1. The solution $x_*(t)$ of the Cauchy problem (1.1) exists on the interval $[0;T]$.

In numerical integration, the values of the solution $x_*(t)$ are usually sought on the grid $\{0 = t_0 < t_1 < t_2 < ... < t_n = T\}$. To focus on the essence of the analysis in this article, only the evenly spaced grid $t_n = nh$ will be considered, where $h = \frac{T}{N}$, $N \in \mathbb{R}^+$.

To construct an approximate solution scheme, we take the Taylor series for the exact solution up to an accuracy of $\mathrm{O}(h^4)$.

$$x_*(t+nh) = x_*(t) + h\dot{x}_*(t) + \frac{h^2}{2}\ddot{x}_*(t) + \frac{h^3}{6}\dddot{x}_*(t) + \frac{h^4}{24}x^{(IV)}{}_*(t) + R_5(t,h), \tag{2.1}$$

where

$$x_*(t) = x_0, \quad \dot{x}_*(t) = f[x_*(t)], \quad \ddot{x}_*(t) = f'[x_*(t)]f[x_*(t)],$$
$$\dddot{x}_*(t) = f''[x_*(t)]f[x_*(t)]f[x_*(t)] + f'[x_*(t)]f'[x_*(t)]f[x_*(t)].$$

The formulation of the main result and its proof is lengthy and involves extensive calculations. To overcome this inconvenient situation, a series of special notations is used in the article. One of these is the interpretation of the function f and its derivatives as operators: $\dot{x}_*(t) = f[x_*(t)]$, $\ddot{x}_*(t) = (f'f)[x_*(t)]$, $\dddot{x}_*(t) = (f''ff + f'f'f)[x_*(t)]$ and so on.

In order shorten the expressions further we introduce the following notation

$$F^0 = f, \quad F^1 = f'f, \quad F^2 = f''ff + f'f'f,$$
$$F^3 = f'''fff + 2f''f'ff + f''ff'f + f'f''ff + f'f'f'f. \tag{2.2}$$

Using the above notation allows us to obtain a compact formula for the derivatives:

$$x^{(k)}(t) = (F^k)[x_*(t)], \text{ for } (k = \overline{0,3}).$$

Therefore, the Taylor expansion of the exact solution have the form

$$x_*[(n+1)h] = x_*(nh) + \sum_{k=1}^{4} \frac{h^k}{k!} F^{k-1}[x_*(nh)] + R_5.$$

If we omit the remainder term in the above formula we obtain the following recurrent formula

$$x_{n+1} = x_n + \sum_{k=1}^{4} \frac{h^k}{k!} F^{k-1}(x_n). \tag{2.3}$$

It is intuitively clear that x_n approximates the value of $x_*(nh)$.

Naturally, in general, x_n does not exactly match $x_*(nh)$, and the issue arises of estimating the difference $|x_*(nh) - x_n|$. For practical applications, an estimate of the form (1.2) is sufficient. However, in theoretical problems, it is also necessary to consider the values of the solution $x_*(t)$ between the node points. Clearly, for $t \in [nh, (n+1)h]$ we can use the inequality

$$|x_*(t) - x_n| \le h \max_{nh \le t \le (n+1)h} |f(x_*(t))|.$$

Since, the order of this approximation is h obtaining higher order estimates for the difference $|x_*(nh) - x_n|$ is meaningless. In order to obtain higher order estimates for $|x_*(t) - \tilde{x}(t)|$ we use the following approximate solution

$$\tilde{x}(t) = x_0 + \int_0^t \sum_{k=0}^{3} \frac{(s-\sigma_s)^k}{k!} F^k[\tilde{x}(\sigma_s)]ds \tag{2.4}$$

where $\sigma_t = h\,[t/h]$ is the step function for the mesh under consideration.

Lemma 2.1. *For every $n \ge 1$ the following equality holds $\tilde{x}(nh) = x_n$.*

It is easily proved by induction on the intervals $[0,\ nh]$.

Assumption 2. Suppose that there exists $K \subset \mathbb{R}^d$ a compact and convex domain such that in the interval $[0;T]$ the solution $x_*(t)$ exists and $\tilde{x}(t) \in K$ for all $t \in [0;T]$.

Introduce the following constants:

$$\begin{aligned} M_0 =& \max_{x\in K} |f(x)|, \quad M_1 = \max_{x\in K} \|f'(x)\|, \quad M_2 = \max_{x\in K} \|f''(x)\| \\ & M_3 = \max_{x\in K} \|f'''(x)\|, \qquad M_4 = \max_{x\in K} \left\|f^{\mathrm{iv}}(x)\right\|. \end{aligned} \tag{2.5}$$

Below the arguments of the functions $f^{(k)}$, $k = 0,1,2,3$ will be either $\tilde{x}(t)$, or $\tilde{x}(\sigma_t)$. In the first case we write the argument completely, while in the second case we shorten it to $f_t^{(k)} = f^{(k)}[x(\sigma_t)]$. Likewise, $F^k[x(\sigma_t)] = F_t^k$, $k = 0,1,2,3$. Further, we introduce the following constants:

$$l_0 = M_0, \ \ l_1 = M_1 M_0, \ \ l_2 = M_2 {M_0}^2 + {M_1}^2 M_0, \ \ l_3 = M_3 {M_0}^3 + 4 M_2 M_1 {M_0}^2 + {M_1}^3 M_0$$

and

$$\begin{aligned} L_3 &= \frac{1}{24} M_1 l_3, \\ L_{2,3} &= \frac{1}{24} M_0 M_2 l_2 + \frac{1}{120} M_0 M_2 l_3 h, \\ L_{1,2,3} &= \frac{1}{8} M_2 {l_1}^2 + \frac{1}{30} M_2 l_1 l_2 h + \frac{1}{144} M_2 l_1 l_3 h^2, \\ {H^1}_{1,2,3} &= \frac{1}{24} {M_0}^2 M_3 l_1 + \frac{1}{120} {M_0}^2 M_3 l_2 h + \frac{1}{720} {M_0}^2 M_3 l_3 h^2, \\ {H^1}_{0,1,2,3} &= \frac{1}{8} {M_2}^2 {M_0}^3 + \frac{1}{20} {M_0}^2 {M_2}^2 l_1 h + \frac{1}{72} {M_0}^2 {M_2}^2 l_2 h^2 + \frac{1}{336} {M_0}^2 {M_2}^2 l_3 h^3, \\ {H^2}_{0,1,2,3} &= \frac{1}{8} M_2 {l_1}^2 + \frac{1}{20} M_1 M_2 {l_1}^2 h + \frac{1}{72} M_1 M_2 l_1 l_2 h^2 + \frac{1}{336} M_1 M_2 l_1 l_3 h^3, \\ {H^3}_{0,1,2,3} &= \frac{1}{12} M_3 {M_0}^2 l_1 + \frac{1}{40} M_0 M_3 {l_1}^2 h + \frac{1}{90} M_0 M_3 l_1 l_2 h^2 + \frac{1}{1008} M_0 M_3 l_1 l_3 h^3, \\ {H^4}_{0,1,2,3} &= \frac{1}{8} M_3 {M_0}^2 l_1 + \frac{1}{30} M_0 M_3 {l_1}^2 h + \frac{1}{144} M_0 M_3 l_1 l_2 h^2 + \frac{1}{840} M_0 M_3 l_1 l_3 h^3, \\ {H^5}_{0,1,2,3} &= \frac{1}{4!} {M_0}^4 M_4 + \frac{1}{5!} {M_0}^3 M_4 l_1 h + \frac{1}{6!} {M_0}^3 M_4 l_2 h^2 + \frac{1}{7!} {M_0}^3 M_4 l_3 h^3. \end{aligned} \tag{2.6}$$

Observe the following inequality

$$\max_{x\in K}\left|F^k(x)\right| \le L_k, \quad k=0,1,2,3.$$

3. The Main Result

To state the main results we need to compute the derivatives of the approximate solution. By the virtue of (2.4) for every $t \neq nh$ and $(k = 1, 2, ..., N-1)$ we have

$$\dot{\tilde{x}}(t) = \sum_{k=0}^{3} \frac{(t-\sigma_t)^k}{k!} {F_t}^k. \tag{3.1}$$

Taking into account $0 \le t - \sigma_t \le h$ we obtain the estimate

$$\left|\dot{\tilde{x}}(t)\right| \le M_0 + l_1 h + l_2 \frac{h^2}{2} + l_3 \frac{h^3}{6}.$$

Theorem 3.1. *Let $x_*(\cdot)$ be the solution of Cauchy problem (1.1) and $\tilde{x}(\cdot)$ be its approximate solution given by (2.4) defined on an interval $[0, T]$. The following inequality holds*

$$|x_*(t) - \tilde{x}(t)| \le Ch^4 \frac{e^{M_1 T} - 1}{M_1},$$

where

$$C = L_3 + L_{2,3} + L_{1,2,3} + {H^1}_{1,2,3} + {H^2}_{0,1,2,3} + {H^3}_{0,1,2,3} + {H^4}_{0,1,2,3} + {H^5}_{0,1,2,3}.$$

Proof: For the exact solution we have $x_*(t) = x_0 + \int\limits_0^t f\left[x_*(s)\right]ds$. Let

$$I(s) = f[\tilde{x}(s)] - \sum_{k=0}^{3} \frac{(s-\sigma_s)^k}{k!} {F_s}^k. \tag{3.2}$$

Using (2.4) and the triangle inequality we obtain

$$|x_*(t) - \tilde{x}(t)| \le \int\limits_0^t |I(s)|ds + \int\limits_0^t |f[x_*(s)] - f[\tilde{x}(s)]|ds. \tag{3.3}$$

We start with estimating the $I(s)$ the aim is to obtain an inequality of the form

$$|I(s)| \le Ch^4.$$

Grouping $f[\tilde{x}(s)]$ with the term $F_S^0 = f_s = f[\tilde{x}(\sigma_s)]$, which corresponds to $k = 0$ in the right hand side of (3.3) and applying the fundamental rule of calculus implies

$$I(s) = \int\limits_{\sigma_s}^{s} \frac{df(\tilde{x}(r))}{dr} dr - \sum_{k=1}^{3} \frac{(s-\sigma_s)^k}{k!} {F_s}^k.$$

Notice that s belongs to $[0, t]$. From now on we fix s. Below all the new variables take values from the interval (σ_s, s). In particular, $\sigma_s \le r < s$ in the expression of $I(s)$, whence, $\sigma_s = \sigma_r$. It implies that $f_s = f_r$. This applies to the variable t as well. Hence, $\sigma_t = \sigma_s$, $f_t = f_s$.

Further, notice that equality (3.1) implies that

$$\dot{\tilde{x}}(r) = \sum_{k=0}^{3} \frac{(r-\sigma_r)^k}{k!} {F_r}^k.$$

In the above equality we denote the term corresponding to $k = 3$ can be written in the form $B_3 = \int\limits_{\sigma_s}^{s} f'[\tilde{x}(r)]\frac{(r-\sigma_r)^3}{3!}{F_r}^3 dr$. Taking into account these last to equalities $I(s)$ can be rewritten as

$$I(s) = \int\limits_{\sigma_s}^{s} f'[\tilde{x}(r)] \sum_{k=0}^{2} \frac{(r-\sigma_r)^k}{k!} {F_r}^k dr - \sum_{k=1}^{3} \frac{(s-\sigma_s)^k}{k!} {F_s}^k + B_3 =$$

$$= \int\limits_{\sigma_s}^{s} \sum_{k=0}^{2} \frac{(r-\sigma_r)^k}{k!} \left[f'[\tilde{x}(r)]{F_r}^k - {F_s}^{k+1} \right] dr + B_3. \tag{3.4}$$

It is clear that for B_3 the inequality

$$|B_3| \leq \frac{(s-\sigma_s)^4}{24} M_1 l_3 \leq \frac{h^4}{24} M_1 l_3 = h^4 L_3 \tag{3.5}$$

holds.

Intuitively it is clear that the terms in (3.4) are estimated exactly by the power of h, which is equal to the number of occurrence of letter f (i.e., f, f',f'', ...). Taking this property into we regroup the term of $I(s)$ as follows:

$$I(s) = I_1(s) + I_2(s) + I_3(s) + B_3,$$

where

$$I_1(s) = \int\limits_{\sigma_s}^{s} \{f'[\tilde{x}(r)]f_r - (f'f)_s\} dr, \tag{3.6}$$

$$I_2(s) = \int\limits_{\sigma_s}^{s} (r-\sigma_r) \left\{ f'[\tilde{x}(r)](f'f)_r - \left((f''ff)_s + (f'f'f)_s \right) \right\} dr, \tag{3.7}$$

$$I_3(s) = \int\limits_{\sigma_s}^{s} \frac{(r-\sigma_r)^2}{2!} \left\{ f'[\tilde{x}(r)]{F_r}^2 - {F_s}^3 \right\} dr. \tag{3.8}$$

We start with the estimation of $I_1(s)$. Using the definition of $\dot{\tilde{x}}(t)$ in equation (3.1) and the equality $f_r = f_s$ we obtain:

$$I_1(s) = \int\limits_{\sigma_s}^{s} \left(\int\limits_{\sigma_r}^{r} \frac{d(f'[\tilde{x}(u)])}{du} du \right) f_r dr = \int\limits_{\sigma_s}^{s} \int\limits_{\sigma_r}^{r} f''[\tilde{x}(u)] \sum_{k=0}^{3} \frac{(u-\sigma_u)^k}{k!} {F_u}^k f_r du dr.$$

Integrating the term corresponding to $k = 2, 3$ twice we obtain terms of the order h^4 and h^5 respectively. Indeed, introducing the following notation:

$$C_{2,3} = \int\limits_{\sigma_s}^{s} \int\limits_{\sigma_r}^{r} f''[\tilde{x}(u)] \sum_{k=2}^{3} \frac{(u-\sigma_u)^k}{k!} {F_u}^k f_r du dr.$$

We have

$$|C_{2,3}| = \left| \int\limits_{\sigma_s}^{s} \int\limits_{\sigma_r}^{r} f''[\tilde{x}(u)] \sum_{k=2}^{3} \frac{(u-\sigma_u)^k}{k!} {F_u}^k f_r du dr \right| \leq h^4 L_{2,3}. \tag{3.9}$$

Using the notation for $C_{2,3}$ expression of $I_1(s)$ can be rewritten in the form

$$I_1(s) = \int\limits_{\sigma_s}^{s} \int\limits_{\sigma_r}^{r} f''[\tilde{x}(u)] \sum_{k=0}^{1} \frac{(u-\sigma_u)^k}{k!} {F_u}^k f_r du dr + C_{2,3}. \tag{3.10}$$

As it is easily seen from (3.10) it is sufficient to obtain estimates on the order of h^4 for the fitst summand. But we delay this until the end. Since we need some other terms coming from I_2, I_3 to group with the terms corresponding to $k=0$ and $k=1$ in (3.10).

Now, we start estimating the term $I_2(s)$. As above, substituting $\dot{\tilde{x}}(t)$ with the expression in (3.1) we obtain:

$$I_2(s)=\int\limits_{\sigma_s}^{s}(r-\sigma_r)\left(\int\limits_{\sigma_s}^{r}f''[\tilde{x}(u)]\sum_{k=0}^{3}\frac{(u-\sigma_u)^k}{k!}{F_u}^k du\right)(f'f)_r dr-\int\limits_{\sigma_s}^{s}(r-\sigma_r)(f''ff)_r dr$$

In this case we denote by $D_{1,2,3}$ the sum of the terms corresponding $k=1,2,3$ of the first term of $I_2(s)$ and obtain

$$|D_{1,2,3}|=\left|\int\limits_{\sigma_s}^{s}(r-\sigma_r)\left(\int\limits_{\sigma_s}^{r}f''[\tilde{x}(u)]\sum_{k=1}^{3}\frac{(u-\sigma_u)^k}{k!}{F_u}^k du\right)(f'f)_r dr\right|\le h^4 L_{1,2,3}. \tag{3.11}$$

Therefore we write

$$I_2(s)=\int\limits_{\sigma_s}^{s}(r-\sigma_r)\left(\int\limits_{\sigma_s}^{r}f''[\tilde{x}(u)]f_u du\right)(f'f)_r dr-\int\limits_{\sigma_s}^{s}(r-\sigma_r)(f''ff)_r dr+D_{1,2,3}. \tag{3.12}$$

Again we leave estimating the remaining terms to future. Finally we apply the same method to $I_3(s)$. First of all using the definition (2.2) we group the elements $f'[\tilde{x}(r)](f''ff)_r$ and $f'[\tilde{x}(r)](f'f'f)_r$ of $f'[\tilde{x}(r)]{F_r}^2$ with the corresponding elements $(f'f''ff)_s$ and $(f'f'f'f)_s$ of ${F_s}^3$ and denote the difference by:

$${E^1}_{0,1,2,3}=\int\limits_{\sigma_s}^{s}\frac{(r-\sigma_r)^2}{2}\left(\int\limits_{\sigma_s}^{r}f''[\tilde{x}(u)]\sum_{k=0}^{3}\frac{(u-\sigma_u)^k}{k!}{F_u}^k du\right)(f''ff)_s dr,$$

$${E^2}_{0,1,2,3}=\int\limits_{\sigma_s}^{s}\frac{(r-\sigma_r)^2}{2!}\left(\int\limits_{\sigma_s}^{r}f''[\tilde{x}(u)]\sum_{k=0}^{3}\frac{(u-\sigma_u)^k}{k!}{F_u}^k du\right)(f'f'f)_s dr.$$

Observe that for ${E^1}_{0,1,2,3}$ and ${E^2}_{0,1,2,3}$ the following estimates hold:

$$\left|{E^1}_{0,1,2,3}\right|\le h^4{H^1}_{0,1,2,3},\quad \left|{E^2}_{0,1,2,3}\right|\le h^4{H^2}_{0,1,2,3}. \tag{3.13}$$

Therefore, for $I_3(s)$ we have

$$I_3(s)={E^1}_{0,1,2,3}+{E^2}_{0,1,2,3}-\int\limits_{\sigma_r}^{s}\frac{(r-\sigma_r)^2}{2!}\left\{(f''f'ff)_s+2(f''f'ff)_s+(f'''fff)_s\right\}dr. \tag{3.14}$$

Now, we start estimating the remaining terms of $I_1(s)$, $I_2(s)$ and $I_3(s)$. We start with groupning the term of (3.10) corresponding to $k=0$ with the second term of (3.12):

$$\int\limits_{\sigma_s}^{s}\int\limits_{\sigma_r}^{r}f''[\tilde{x}(u)]f_r f_u du dr-\int\limits_{\sigma_s}^{s}(r-\sigma_r)(f''ff)_r dr=\int\limits_{\sigma_s}^{s}\int\limits_{\sigma_r}^{r}\int\limits_{\sigma_r}^{u}f'''[\tilde{x}(v)]\sum_{k=0}^{3}\frac{(v-\sigma_v)^k}{k!}{F_v}^k(ff)_r dv du dr.$$

Now, denoting by $E_{1,2,3}$ the sum of terms corresponding to $k=1,2,3$ under the integral in the above formula we obtain

$$|E_{1,2,3}|=\left|\int\limits_{\sigma_s}^{s}\int\limits_{\sigma_r}^{r}\int\limits_{\sigma_r}^{u}f'''[\tilde{x}(v)]\sum_{k=1}^{3}\frac{(v-\sigma_v)^k}{k!}{F_v}^k(ff)_r dv du dr\right|\le h^4{H^1}_{1,2,3}. \tag{3.15}$$

Substituting the inequalities (3.10), (3.12) and (3.14) into $I(s)$ we obtain

$$I(s)=\int_{\sigma_s}^{s}\int_{\sigma_r}^{r}(u-\sigma_u)f''[\tilde{x}(u)]f_r(f'f)_u dudr-\int_{\sigma_r}^{s}\frac{(r-\sigma_r)^2}{2!}(f''f'ff)_s dr+$$
$$+\int_{\sigma_s}^{s}(r-\sigma_r)\left(\int_{\sigma_s}^{s}f'''[\tilde{x}(u)]f_u du\right)(f'f)_r dr-\int_{\sigma_r}^{s}\frac{(r-\sigma_r)^2}{2!}(2f''f'ff)_s dr+$$
$$+\int_{\sigma_s}^{s}\int_{\sigma_r}^{r}\int_{\sigma_r}^{u}f'''[\tilde{x}(v)]f_v(ff)_r dvdudr-\int_{\sigma_r}^{s}\frac{(r-\sigma_r)^2}{2!}(f'''fff)_r dr+ \tag{3.16}$$
$$+B_3+C_{2,3}+D_{1,2,3}+E_{1,2,3}+{E^1}_{0,1,2,3}+{E^2}_{0,1,2,3}$$

In equation (3.16) we couple the first and second terms and apply the fundamental rule of calculus, then use equation (3.1) and obtain the following equation:

$${E^3}_{0,1,2,3}=\int_{\sigma_s}^{s}\int_{\sigma_r}^{r}(u-\sigma_u)\left(\int_{\sigma_s}^{u}f'''[\tilde{x}(v)]\sum_{k=0}^{3}\frac{(v-\sigma_v)^k}{k!}{F_v}^k dv\right)(f'ff)_s dudr.$$

In the same way, by coupling the first term with the fourth, and the fifth term with the sixes we obtain

$${E^4}_{0,1,2,3}=\int_{\sigma_s}^{s}(r-\sigma_r)\left(\int_{\sigma_s}^{s}\left(\int_{\sigma_s}^{u}f'''[\tilde{x}(v)]\sum_{k=0}^{3}\frac{(v-\sigma_v)^k}{k!}{F_v}^k dv\right)(f'ff)_r du\right)dr,$$
$${E^5}_{0,1,2,3}=\int_{\sigma_s}^{s}\int_{\sigma_r}^{r}\int_{\sigma_r}^{u}\left(\int_{\sigma_s}^{v}f^{(IV)}[\tilde{x}(z)]\sum_{k=0}^{3}\frac{(z-\sigma_z)^k}{k!}{F_z}^k dz\right)(fff)_r dvdudr.$$

The following inequalities hold:

$$\left|{E^3}_{0,1,2,3}\right|\le h^4{H^3}_{0,1,2,3},\quad \left|{E^4}_{0,1,2,3}\right|\le h^4{H^4}_{0,1,2,3},\quad \left|{E^5}_{0,1,2,3}\right|\le h^4{H^5}_{0,1,2,3}. \tag{3.17}$$

Using the above notations we rewrite $I(s)$ in the form:

$$I(s)=B_3+C_{2,3}+D_{1,2,3}+E_{1,2,3}+{E^1}_{0,1,2,3}+{E^2}_{0,1,2,3}+{E^3}_{0,1,2,3}+{E^4}_{0,1,2,3}+{E^5}_{0,1,2,3}. \tag{3.18}$$

Taking into account the equations (3.18) (3.5), (3.9), (3.11), (3.13), (3.15) and (3.17) from the above equality we deduce the following inequality:

$$|I(s)|\le Ch^4.$$

Since K is a compact set by assumption, we can use the Mean Value theorem (Hadamard lemma) to estimate the difference $f[x_*(s)]-f[\tilde{x}(s)]$. We have

$$|f[x_*(s)]-f[\tilde{x}(s)]|\le M_1\,|x_*(s)-\tilde{x}(s)|\,.$$

Therefore,

$$|x_*(t)-\tilde{x}(t)|\le\int_0^t|I(s)|ds+\int_0^t|f[x_*(s)]-f[\tilde{x}(s)]|ds\le Ch^4t+\int_0^t M_1\,|x_*(s)-\bar{x}(s)|ds. \tag{3.19}$$

We apply the following form of the Grönwall inequality: Suppose that $u(t)\ge 0$ is a continuous function for $t\ge 0$, and C,M are non-negative numbers such that $u(t)\le Ct+M\int_0^t u(s)ds$. Then the following estimate holds:

$$u(t)\le C\frac{e^{Mt}-1}{M}.$$

Using the above inequality in (3.19) implies

$$|x_*(t) - \tilde{x}(t)| \leq Ch^4 \frac{e^{M_1 T} - 1}{M_1},$$

which finishes the proof.

□

4. Applications to Van der Pol oscillator

Here we apply the results to the Van der Pol oscillator:

$$\begin{cases} \dot{x} = y \\ \dot{y} = \mu(1 - x^2)y - x \end{cases} \tag{4.1}$$

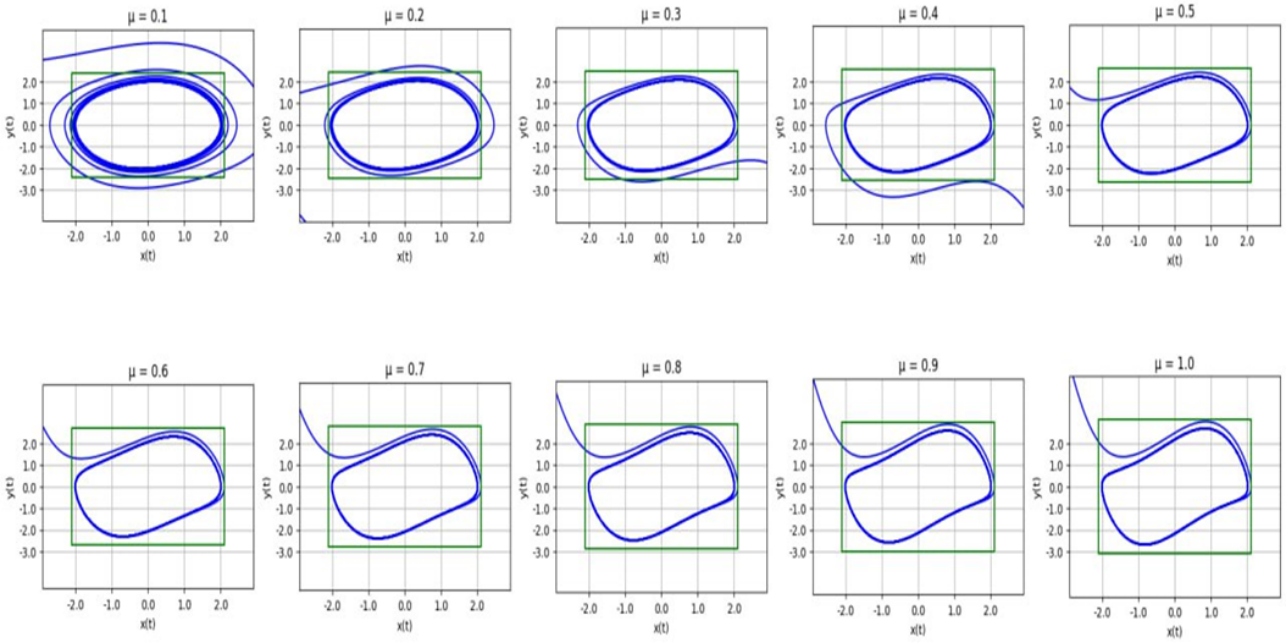
to approximate the period of its periodic trajectories for the parameters $\mu = i/10$, $i = 1, ..., 10$ and compare it with the known formula for the periods:

$$T = \frac{2\pi}{1 - \frac{\mu^2}{16} + \frac{17\mu^4}{3072} + O(\mu^6)}$$

in the rectangle $K = [(x, y) \in R \,||x| \leq 2.1, \;\; |y| \leq MaxY + 0.001]$

Table 1. Numerical results for different values of μ

μ	Max x	Max y	M_0	M_1	M_2	M_3	M_4	T	Error
0.1	2.0001	2.0116	3.44491	2.12964	0.718523	0.34641	0	6.2871	$2.007251\ 10^{-52}$
0.2	2.0004	2.0442	4.05912	2.98222	1.44442	0.69282	0	6.2988	$1.188056\ 10^{-48}$
0.3	2.0009	2.0935	4.74375	3.92072	2.1836	1.03923	0	6.3082	$6.209675\ 10^{-45}$
0.4	2.0016	2.1554	5.49799	4.93663	2.94036	1.38564	0	6.3457	$2.974666\ 10^{-41}$
0.5	2.0025	2.2270	6.32302	6.03082	3.71809	1.73205	0	6.3807	$1.327857\ 10^{-37}$
0.6	2.0035	2.3063	7.22075	7.20637	4.5196	2.07846	0	6.4231	$5.594606\ 10^{-34}$
0.7	2.0047	2.3919	8.19331	8.46681	5.3474	2.42487	0	6.4728	$2.247712\ 10^{-30}$
0.8	2.0059	2.4830	9.24344	9.81633	6.20401	2.77128	0	6.5296	$8.680078\ 10^{-27}$
0.9	2.0072	2.5787	10.3729	11.2579	7.09149	3.11769	0	6.5930	$3.242253\ 10^{-23}$
1.0	2.0086	2.6784	11.5837	12.7947	8.01186	3.4641	0	6.6627	$1.177264\ 10^{-19}$



5. Acknowledgment

This research was carried out under Research Project No. FL-9524115148 of the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan.

References

[1] Bakhvalov N. S., Numerical Methods. Nauka, Moscow, (1973).

[2] Berezin B. C., Jidkov N. N., Metodi vichisleniy. Moskva, (1959).

[3] Butcher J. C., Numerical Methods for Ordinary Differential Equations. John Wiley Sons, Ltd., United Kingdom, (2016).

[4] Hairer E., Norsett S., Wanner G., Solving ordinary differential equations. I. (Nonstil problems). Springer, (2008).

[5] Lambert J. D., Numerical methods for ordinary differential systems. The initial value problems. J. Wiley Sons, (1993).

[6] Buchberger B., Collins G. E., Loos R., Computer Algebra: Symbolic and Algebraic Computation. Springer Science Business Media, (2013).

[7] Gathen J., Gerhard J., Modern Computer Algebra. Cambridge University Press, (2013).

[8] Geddes K. O., Czapor S. R., Labahn G., Algorithms for Computer Algebra. Springer Science Business Media, (2007).

[9] Guckenheimer J., Numerical analysis of dynamical systems. Elsevier, (2002).

[10] Azamov A. A., Akhmedov O. S., Existence of a complex closed trajectory in a three-dimensional dynamical system. Comput. Math. and Math. Phys. (2011) Vol. 51, P. 1353–1359.

[11] Azamov A., Abdullayev A., Ruziboev M., An explicit estimate for approximate solutions of ODEs based on the Taylor formula. Ural Mathematical Journal, (2024) Vol. 10, Iss. 1, P. 18–27.

[12] Kamke E., Handbook of Ordinary Differential Equations. Nauka, (1971).

[13] Kendall A. E., An Introduction to Numerical Analysis. John Wiley Sons, Inc., (1989).

[14] Babenko K. I., Foundations of Numerical Analysis [Osnovy chislennogo analiza]. Regulyarnaya i haoticheskaya dinamika, Moscow, Izhevck, (2002).

[15] Krilov V. I., Bobkov V. V., Monastirniy P. I., Vichislitelnie metodi. Nauka, Moskva, (1977).

[16] Tucker W., The Lorenz attractor exists. Compt. Rend. de l'Académie des Sciences - Series I - Mathematics, (1999) Vol. 328, Iss. 12, P. 1197–1202.

[17] Tucker W., A Rigorous ODE Solver and Smale's 14th Problem. Found. Comput. Math. (2002) Vol. 2, P. 53–117.

[18] Berg J. B., Sheombarsing R., Rigorous numerics for ODEs using Chebyshev series and domain decomposition. Journal of Computational Dynamics, (2021) Vol. 8, Iss. 2158-2505, P. 353.

[19] Gameiro M., Lessard J.-P., Rigorous computation of smooth branches of equilibria for the three dimensional Cahn-Hilliard equation. Numer. Math. (2011) Vol. 117, Iss. 4, P. 753–778.

[20] Berg J. B., Mireles-James J. D., Lessard J.-P., Mischaikow K., Rigorous numerics for symmetric connecting orbits: even homoclinics of the Gray-Scott equation. SIAM J. Math. Anal. (2011) Vol. 43, Iss. 4, P. 1557–1594.

[21] Hartman P., Ordinary differential equations. SIAM, (2002).

Abdullaev A.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan
e-mail: abduganiax@mail.ru

Azamov A.A.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan
e-mail: abdulla.azamov@gmail.com

Ruziboev M.B.,
National Pedagogical University of Uzbekistan,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan
e-mail: m.ruziboev@npuu.uz

Tilavov A.M.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan;
e-mail: asliddintm@mail.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 15–23
DOI: 10.29229/uzmj.2026-2-2

Schrödinger equation is a weak form of the classical Euler-Lagrange equation

Akrami S. E.

Abstract. Inspired by quantum mechanics, we introduce a weak form of solutions for differential equations. We show that Schrödinger equation is a weak form of the classical Euler-Lagrange equation.

Keywords: partial differential equations, weak solution, quantum mechanics, Schrödinger equation, Euler-Lagrange equation
MSC (2020): 35D30, 35Q40, 70H03, 58D30, 81Q05, 49S05

1. Introduction

For partial differential equations various forms of weak solutions are defined, [1]. In this paper, we define a new form of weak solutions for differential equations. Motivation comes from quantum mechanics where Newton equation is replaced with the Schrödinger equation. In classical mechanics, Newton equation for a particle is

$$m\ddot{x}(t) + \nabla U(x(t)) = 0,$$

where m is a positive constant, $x(t)$ is a curve in $\mathbb{R}^3$ and $U : \mathbb{R}^3 \to \mathbb{R}$. In quantum mechanics this equation is replaced with the Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla\psi + U\psi, \tag{1.1}$$

where $\psi(t,x) \in \mathbb{C}, t \in \mathbb{R}, x \in \mathbb{R}^3$. There is a big radical change in passing from Newton equation to Schrödinger equation. We try to bring these two different equations close to each other. To do this task, we first rewrite Newton equation as follows

$$\dot{x}(t) = v(t), \tag{1.2}$$

and

$$m\dot{v}(t) + \nabla U(x(t)) = 0. \tag{1.3}$$

The first equation just means that the variable $v(t)$ is the derivative of $x(t)$. Next, using the polar decomposing of the complex function $\psi(t,x) = R(t,x)e^{\frac{i}{\hbar}S(t,x)}$ and setting $\rho(t,x) := R^2(t,x), V(t,x) := \frac{1}{m}\nabla S(t,x)$ we rewrite the Schrödinger equation as follows

$$\frac{\partial\rho}{\partial t} + \nabla.(\rho V) = 0, \tag{1.4}$$

and

$$m\Big(\frac{\partial V}{\partial t} + (V\cdot\nabla)V\Big) + \nabla U + \nabla Q = 0,$$

where the term Q is called quantum potential and it is known that $\int \rho\nabla Q dx = 0$, [2]. Thus

$$\int \rho\Big(m(\frac{\partial V}{\partial t} + (V\cdot\nabla)V) + \nabla U\Big)dx = 0. \tag{1.5}$$

We call equations (1.4) and (1.5) as *weak form* of equations (1.2) and (1.3). This observation suggests to replace a curve $x(t) \in \mathbb{R}^3$ and its derivative $v(t) = dx/dt$ by probability density $\rho(t,x)$ and vector field $V(t,x)$ satisfying (1.4). We will call the pair $(x(t), v(t))$ as a *strong differentiable function* and the pair $(\rho(t,x), V(t,x))$ as a *weak differentiable function*.

Ambrosio et al., [3], have studied differential calculus over a metric measure space and in particular over Wasserstein space of probability measures. We briefly recall this theory. First, let (S, d) be a complete metric space. For any curve $\rho : (a, b) \subseteq \mathbb{R} \to S$ the limit

$$|\rho'|(t) := \lim_{h\to 0} \frac{d(\rho(t+h), \rho(t))}{|h|},$$

if exists, is called the **metric derivative** of ρ. For any absolutely continuous curve $\rho : (a, b) \subseteq \mathbb{R} \to S$ the metric derivative exists for Lebesgue-a.e. $t \in (a, b)$ and we have $d(\rho(s), \rho(t)) \leq \int_s^t |\rho'|(r)dr$ for any interval $(s, t) \subseteq (a, b)$. In the case where S is a Banach space, a curve $\rho : (a, b) \to S$ is absolutely continuous if and only it is differentiable in the ordinary sense for Lebesgue-a.e. $t \in (a, b)$ and we have $\|\rho'(t)\| = |\rho'|(t)$ for Lebesgue-a.e. $t \in (a, b)$. Next in [3], this fact has been applied to the metric space of probability measures on a Hilbert space X. This space is equipped with the Wasserstein's metric. It is shown that the class of absolutely continuous curves ρ_t in this metric space coincides with solutions of the continuity equation of physicists. More precisely, given an absolutely continuous curve ρ_t, one can find a Borel time-dependent velocity field $V_t : X \to X$ such that $\|V_t\|_{L^p(\rho_t)} \leq |\rho'|(t)$ for a.e. t and the continuity equation $\frac{\partial \rho}{\partial t} + \nabla.(\rho V) = 0$ holds. Conversely, if ρ_t solves the continuity equation for some Borel velocity field V_t with $\int_a^b \|V_t\|_{L^p(\rho_t)} dt < \infty$, then ρ_t is an absolutely continuous curve and $\|V_t\|_{L^p(\rho_t)} \leq |\rho'|(t)$ for a.e. $t \in (a, b)$. As a consequence, we see that among all velocity fields V_t which produce the same flow ρ_t, there is a unique optimal one with the smallest $L^p(\rho_t, X)$-norm, equal to the metric derivative of ρ_t. One can view this optimal field as the "tangent" vector field to the curve ρ_t. Next, Villani, Lott, and Sturm studied the concept of Ricci curvature for a metric measure space [4, 5]. The references [6, 7, 8, 9, 4] studied Riemannian geometry and dynamics on the Wasserstein space by putting the structure of an informal manifold, first introduced by Otto [10].

In this paper, we will work with the informal differentiable structure on the Wasserstein space. In Section 2, we will define the concept of a weak differentiable function. In Section 3, we will state the weak form of the Euler-Lagrange equation and derive Schrödinger equation.

2. Weakly Differentiable Functions

This paper is based on the following observation which redefines the Newton-Leibniz derivative of a function.

Theorem 2.1. *A curve $x(t) \in M$ over a manifold M is differentiable if and only if there exists a vector field $v(t) \in T_{x(t)}M$ along $x(t)$ such that*

$$\frac{d}{dt} f(x(t)) = df_{x(t)}(v(t)),$$

for all differentiable functions f on M and moreover we have $v(t) = dx(t)/dt$.

In this paper, we call the function $x(t) \in M$ a *strong differentiable curve.*

Definition 2.2. A **weak curve** over an oriented Riemannian manifold M is a curve ρ_t of probability densities over M, i.e. $\rho(t, x) \geq 0, a \leq t \leq b, x \in M$, $\int_M \rho(t, x)dx = 1$ for all $a \leq t \leq b$, and ρ vanishes at infinity with respect to the x-variable, where integral is with respect to the Riemannian volume form of M.

Next, we define the *derivative of a weak curve* by imitating Theorem 2.1.

Definition 2.3. A weak curve ρ over M is called differentiable if there exists a differentiable time-dependent vector field V_t over M such that

$$\frac{d}{dt} \int \rho(t, x) f(x)dx = \int \rho(t, x)df(V(t, x)), \tag{2.1}$$

for all $f \in C_c^\infty(M)$, in the space of all smooth compactly-supported functions on M.

Theorem 2.4. *A weak curve ρ is differentiable if and only if there exists a differentiable time-dependent vector field V_t such that*

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho V) = 0. \tag{2.2}$$

Proof: We have $\frac{d}{dt}\int \rho f = \int \frac{\partial \rho}{\partial t} f$. On the other hand by the divergence theorem and that f has compact support and ρ vanishes at infinity, we have

$$\int \rho df(V) = \int \nabla.(\rho f V) - \int f \nabla.(\rho V) = -\int f \nabla.(\rho V).$$

Thus (2.1) holds if and only if $\int \Big(\frac{\partial \rho}{\partial t} + \nabla.(\rho V)\Big) f = 0, \forall f$. The later is equivalent with (2.2). □

Note that the vector field V is not unique. Thus we redefine our concept of differentiability as follows.

Definition 2.5. A **differentiable weak curve** is a pair (ρ_t, V_t) where ρ_t is a curve of probability densities and V_t is a vector field over M such that (2.2) holds.

Next, we define a several variable differentiable weak function.

Definition 2.6. A *differentiable weak function* from an open subset $U \subseteq \mathbb{R}^m$ into M is a system $(\rho, V_1, \cdots, V_m)$ including a differentiable function $\rho : U \times M \to [0, \infty)$, vanishing at infinity with respect to M, together with differentiable vector fields $V_j : U \times M \to TM$, $1 \leq j \leq m$, such that

$$\frac{\partial \rho}{\partial u_j} + \nabla.(\rho V_j) = 0.$$

A weak function $\rho : U \times M \to [0, \infty)$ is a replacement for a strong function $f : U \to M$ and the vector fields V_i are replacements for the partial derivatives $V_i = \frac{\partial f}{\partial u_i}$. For the strong function f we have $\frac{\partial^2 f}{\partial u_i \partial u_j} = \frac{\partial^2 f}{\partial u_j \partial u_i}$. Namely, for $V_i = \frac{\partial f}{\partial u_i}$ we have

$$\frac{\partial V_i}{\partial u_j} = \frac{\partial V_j}{\partial u_i}.$$

For weak functions, we have the following result.

Theorem 2.7. *If $(\rho, V_1, \cdots, V_m)$ is a differentiable weak function from an open region $U \subseteq \mathbb{R}^m$ to M then*

$$\rho\Big(\frac{\partial V_i}{\partial u_j} - \frac{\partial V_j}{\partial u_i} - [V_i, V_j]\Big) = 0.$$

Here $[V_i, V_j]$ means the Lie bracket of vector fields V_i and V_j with respect to M.

Proof: For any compactly-supported function g on M, since ρ vanishes at infinity, by the divergence theorem we have

$$\frac{\partial}{\partial u_i}\int \rho g = \int \frac{\partial \rho}{\partial u_i} g = -\int g \nabla.(\rho V_i) = -\int \nabla(\rho g V_i) + \int \rho V_i g = \int \rho V_i g. \tag{2.3}$$

Thus

$$\begin{aligned} \frac{\partial^2}{\partial u_j \partial u_i}\int \rho g &= \int \frac{\partial \rho}{\partial u_j} V_i g + \int \rho \frac{\partial V_i}{\partial u_j} g = -\int \nabla.(\rho V_j) V_i g + \int \rho \frac{\partial V_i}{\partial u_j} g \\ &= -\int \nabla.(\rho (V_i g) V_j) + \int \rho V_j V_i g + \int \rho \frac{\partial V_i}{\partial u_j} g = \int \rho (\frac{\partial V_i}{\partial u_j} + V_j V_i) g. \end{aligned}$$

Hence

$$0 = \frac{\partial^2}{\partial u_j \partial u_i}\int \rho g - \frac{\partial^2}{\partial u_i \partial u_j}\int \rho g = \int \rho (\frac{\partial V_i}{\partial u_j} + V_j V_i - \frac{\partial V_j}{\partial u_i} - V_i V_j) g = \int \rho (\frac{\partial V_i}{\partial u_j} - \frac{\partial V_j}{\partial u_i} + [V_i, V_j]) g.$$

Since g is arbitrary, the desired identity is obtained. □

Remark 2.8. Without using the divergence theorem, we can prove at most the following identity

$$\nabla.\Big(\rho(\frac{\partial V_i}{\partial u_j}-\frac{\partial V_j}{\partial u_i}-[V_i,V_j])\Big)=0.$$

To prove the later, first, we state an important identity that we discovered recently and we have not seen before elsewhere; one can show by some calculation that for any two vector fields V and W and any function f on M we have

$$\nabla.\Big(\nabla.(fW)V\Big)-\nabla.\Big(\nabla.(fV)W\Big)=\nabla.(f[W,V]).$$

Now using this identity we have

$$\frac{\partial^2\rho}{\partial u_j\partial u_i}=-\frac{\partial}{\partial u_j}\nabla.(\rho V_i)=-\nabla.(\frac{\partial\rho}{\partial u_j}V_i)-\nabla.(\rho\frac{\partial V_i}{\partial u_j})=\nabla.\Big(\nabla.(\rho V_j)V_i\Big)-\nabla.(\rho\frac{\partial V_i}{\partial u_j}).$$

Thus

$$\begin{aligned}
0 &= \frac{\partial^2\rho}{\partial u_j\partial u_i}-\frac{\partial^2\rho}{\partial u_i\partial u_j}\\
&= \nabla.\Big(\nabla.(\rho V_j)V_i\Big)-\nabla.\Big(\nabla.(\rho V_i)V_j\Big)-\nabla.(\rho\frac{\partial V_i}{\partial u_j})+\nabla.(\rho\frac{\partial V_j}{\partial u_i})\\
&= -\nabla.(\rho[V_i,V_j])-\nabla.(\rho\frac{\partial V_i}{\partial u_j})+\nabla.(\rho\frac{\partial V_j}{\partial u_i})\\
&= \nabla.\Big(\rho(\frac{\partial V_j}{\partial u_i}-\frac{\partial V_i}{\partial u_j}-[V_i,V_j])\Big).
\end{aligned}$$

Example 2.9. Let $A\in M_{n\times m}(\mathbb{R})$ be a matrix of size $n\times m$ with real entries whose columns are $A_i, 1\le i\le m$, and let σ be a probability density over $\mathbb{R}^n$. Then the pair $(\rho,V_1,\cdots,V_m)$ given below is a differentiable weak function from $\mathbb{R}^m$ into $\mathbb{R}^n$.

$$\rho(x,y):=\sigma(y-Ax),\qquad V_i(x,y):=A_i.$$

Proof: We have $\frac{\partial\rho}{\partial x_i}=-\nabla\sigma(y-Ax).A_i$ and $\nabla^y.(\rho(x,y)V_i(x,y))=\nabla^y.(\sigma(y-Ax)A_i)=\nabla\sigma(y-Ax).A_i$. □

Definition 2.10. A *differentiable weak function* from a manifold P into an oriented Riemannian manifold M is a pair (ρ,V) including a differentiable function $\rho:P\times M\to[0,\infty)$, together with a differentiable map

$$V:TP\times M\to TM,\qquad V(w,q)\in T_qM,$$

$\forall w\in TP, q\in M$, such that V is linear concerning the first variable and for any coordinate system $u=(u_1,\cdots,u_m)\in U\subseteq\mathbb{R}^m$ for P

$$\frac{\partial\rho}{\partial u_j}+\nabla.(\rho V_j)=0, \tag{2.4}$$

where $V_j:U\times M\to TM, V_j(u,q):=V(\frac{\partial}{\partial u_j}(u),q)\in T_qM$. Here the divergence operator $\nabla.$ is relative to the variable $q\in M$.

Proposition 2.11. *Equations (2.4) do not depend on the coordinate systems of P.*

Proof: Let $W_i(v):=V_p(\frac{\partial}{\partial v_i}(p))$ where $(v_1,\cdots,v_m)$ is another coordinate system for P. Then since $\frac{\partial}{\partial v_i}=\sum_j\frac{\partial u_j}{\partial v_i}\frac{\partial}{\partial u_j}$ we get $W_i(v)=\sum_j\frac{\partial u_j}{\partial v_i}V_j(u)$. Thus

$$\begin{aligned}
&\frac{\partial\rho(v,q)}{\partial v_i}+\nabla.(\rho(v,q)W_i(v,q))\\
= &\sum_j\frac{\partial u_j}{\partial v_i}\frac{\partial\rho(u,q)}{\partial u_j}+\nabla.(\rho(u,q)\sum_j\frac{\partial u_j}{\partial v_i}V_j(u,q))\\
= &\sum_j\frac{\partial u_j}{\partial v_i}[\frac{\partial\rho(u,q)}{\partial u_j}+\nabla.(\rho(u,q)V_j(u,q))].
\end{aligned}$$

Thus since the matrix $(\frac{\partial u_j}{\partial v_i})$ is invertible we conclude that the expression $\frac{\partial \rho(v,q)}{\partial v_i} + \nabla.(\rho(v,q)W_i(v,q))$ vanishes if and only if $\frac{\partial \rho(u,q)}{\partial u_i} + \nabla.(\rho(u,q)V_i(u,q))$ vanishes.

□

3. Weak Form of Euler-Lagrange Equation

Classical mechanics over Wasserstein space of probability measures over a manifold has been studied extensively, [6, 7, 11, 12]. In this section, we are going to give another viewpoint on this subject. In classical mechanics over strong states, i.e. points over a finite-dimensional configuration manifold M, we study the following action functional

$$S(\alpha) = \int_a^b L(\alpha(t), \dot{\alpha}(t))dt, \tag{3.1}$$

over all paths $\alpha(t) \in M, a \leq t \leq b$, and $L : TM \to \mathbb{R}$ is a Lagrangian. The Lagrangian can have an arbitrary, i.e. non-restrictive, dependence to both α and its derivative. Now, in this paper we are going to replace the space M with the space $W^\infty(M)$ of smooth probability measures over M, i.e. smooth Wasserstein space. The first natural question would be what is the correct counterpart of the above action functional in the case of the space $W^\infty(M)$? The Lagrangian in the above action is a function of α and its derivative. So the Lagrangian over $W^\infty(M)$ must be a function of a path $d\mu_t = \rho(t)dx \in W^\infty(M)$ and its derivative. In this paper we work with probability measures which have density with respect to the volume form of Riemannian manifold M where its Riemannian volume form is denoted by dx. We know that the time-derivative of $\rho(t)$ is a vector field $V(t)$ over M satisfying the continuity equation in the sense of metric measure space theory. So the Lagrangian must be a function of ρ and V. The first candidate for an action functional over $W^\infty(M)$ is

$$S(\rho, V) = \int_a^b \int_M \rho(t,x)L(x, V(t,x))dxdt,$$

where $L : TM \to \mathbb{R}$ is a classical Lagrangian over M and dx is the Riemannian volume form of M. But this is too restrictive Lagrangian since its dependence to the density ρ is very simple unlike the action (3.1) which has nonrestrictive dependence to α. Thus we must add some extra term to the Lagrangian which depends to ρ and perhaps to its spatial derivatives.

In [7] the action functional is of the form

$$S(\mu, V) = \frac{1}{2}\int_a^b \int_{\mathbb{R}^n} \|V(t,x)\|^2 d\mu_t(x)dt - \int_a^b U(\mu_t)dt,$$

where μ_t is a path of probability measures and V_t is vector field over $\mathbb{R}^n$. Here, we do not assume that the continuity equation between μ_t and V_t holds.

In an unpublished work[a] the action functional is of the following form

$$S(\mu, V) = \int_a^b \int_{\mathbb{R}^n} L(x, V(t,x))d\mu_t(x)dt + \int_a^b F(\mu_t)dt + G(\mu(b)),$$

where μ_t is a curve of probability measures over $\mathbb{R}^n, V_t$ is a curve of vector fields over $\mathbb{R}^n, L : T\mathbb{R}^n \to \mathbb{R}$, and F as well as G are functionals over the space of probability measures and the continuity equation between μ_t and V_t holds, i.e. $\frac{\partial \mu_t}{\partial t} + \nabla.(\mu_t V_t) = 0$ in the sense of distributions.

In this section, we study the following action functional

$$S(\rho, V) = \int_a^b \int_M \rho(t,x)\Big(L(x, V(t,x)) + F(\rho, \partial_i\rho, \partial^2_{ij}\rho, \partial^3_{ijk}\rho, \cdots)\Big)dxdt, \tag{3.2}$$

for some function $F(y, y_i, y_{ij}, y_{ijk}, \cdots)$ and where $\partial_i = \frac{\partial}{\partial x_i}, \partial^2_{ij} = \frac{\partial}{\partial x_i \partial x_j}$, etc. Here, we assume that (ρ, V) is a weak differentiable function.

[a] P. Cardaliaguet, Paris Dauphine University, April 2019: Analysis in the space of measures. Lecture notes `https://dottorato.math.unipd.it/sites/default/files/Pierre_Cardaliaguet.pdf`

For simplicity, we will work on space $M = \mathbb{R}^n$ but the result can be extended to an arbitrary manifold. In order to be able to apply the least action principle to the action 3.2, we first need to define a perturbation of the differentiable curve (ρ_t, V_t). To do this task, we recall the meaning of a perturbation of a strong differentiable curve $x(t) \in M, a \leq t \leq b$. It is just a two-variable function $x(t,s) \in M, a \leq t \leq b, -\epsilon \leq s \leq \epsilon$, such that $x(t,0) = x(t), x(a,s) = x(a)$ and $x(b,s) = x(b)$. Notice, that we have the following identity

$$\frac{\partial^2 x}{\partial s \partial t} = \frac{\partial^2 x}{\partial t \partial s}.$$

We define a variation of a differentiable weak curve $(\rho_t, V_t), a \leq t \leq b$, over M to be a triple $(\rho_{t,s}(x), V_{t,s}(x), W_{t,s}(x))$ for some parameter s in a small interval around zero, satisfying

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho V) = 0, \qquad \frac{\partial \rho}{\partial s} + \nabla.(\rho W) = 0,$$

and moreover for $s = 0$ we recover the original pair (ρ_t, V_t) and also $\rho_{a,s}(x) = \rho_a(x), \rho_{b,s}(x) = \rho_b(x), W_{a,s}(x) = W_{b,s}(x) = 0, \forall x$. The notations in the following lemma will be used in our calculus of variation.

Lemma 3.1. *Let A and B be two vectors in $\mathbb{R}^n$ and let X be an $n \times n$ matrix. Denote by AB the inner product and by XA the vector XA^T. Then we have*
i) $(XA)B = (XB)A$,
ii) $(AX)B = A(XB)$.

The proof is obvious.

Theorem 3.2. *By the variation of the action functional (3.2) under the above meaning of variation we arrive to the following equation which we call* weak Euler-Lagrange Equation

$$\rho\Big((\frac{\partial}{\partial t} + V\frac{\partial}{\partial x})\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} - \frac{\partial}{\partial x}F - \frac{\partial}{\partial x}(\rho\frac{\partial F}{\partial y} - \frac{\partial}{\partial x_i}(\rho\frac{\partial F}{\partial y_i}) + \frac{\partial^2}{\partial x_i \partial x_j}(\rho\frac{\partial F}{\partial y_{ij}}) - \cdots)\Big) = 0, \tag{3.3}$$

where ∇ is the gradient operator with respect to the variable $x \in M$ and in general for given vector fields X and function f, by $(X.\nabla)f$ we mean the derivative of function f in the direction of vector field X. Here for repeated indexes we use Einstein summation rule. In particular, if the following identity holds for all function $\rho(x)$

$$\rho\frac{\partial F}{\partial y} - \frac{\partial}{\partial x_i}(\rho\frac{\partial F}{\partial y_i}) + \frac{\partial^2}{\partial x_i \partial x_j}(\rho\frac{\partial F}{\partial y_{ij}}) - \cdots = 0, \tag{3.4}$$

then the weak Euler-Lagrange equation becomes

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho V) = 0, \quad \rho\Big((\frac{\partial}{\partial t} + V.\nabla)\frac{\partial L}{\partial \dot{x}}(x,V) - \frac{\partial L}{\partial x}(x,V) - \frac{\partial}{\partial x}F\Big) = 0. \tag{3.5}$$

Proof: We set

$$S_1 := \int_a^b \int_M \rho L(x,V) dx dt,$$

and

$$S_2 := \int_a^b \int_M \rho F(\rho, \partial_i \rho, \partial^2_{ij}\rho, \partial^3_{ijk}\rho, \cdots) dx dt.$$

Let $(\rho_{t,s}(x), V_{t,s}(x), W_{t,s}(x))$ be a variation of the pair (ρ_t, V_t). We set $K := L(x, V(t,s,x))$. Based on the notation introduced in the previous Lemma, we compute the derivative

$$\begin{aligned}
\frac{dS_1}{ds} &= \iint \Big(\frac{\partial \rho}{\partial s}K + \rho\frac{\partial L}{\partial \dot{x}}\frac{\partial V}{\partial s}\Big)dxdt \\
&= \iint \rho\Big(\frac{\partial K}{\partial x}W + \frac{\partial L}{\partial \dot{x}}\frac{\partial V}{\partial s}\Big)dxdt \\
&= \iint \rho\Big(\frac{\partial K}{\partial x}W + \frac{\partial L}{\partial \dot{x}}\frac{\partial W}{\partial t} + \frac{\partial L}{\partial \dot{x}}[V,W]\Big)dxdt \\
&= \iint \Big(\rho\frac{\partial K}{\partial x}W + \frac{\partial}{\partial t}(\rho\frac{\partial L}{\partial \dot{x}}W) - \rho(\frac{\partial}{\partial t}\frac{\partial L}{\partial \dot{x}})W - \frac{\partial \rho}{\partial t}\frac{\partial L}{\partial \dot{x}}W + \rho\frac{\partial L}{\partial \dot{x}}[V,W]\Big)dxdt \\
&= \iint \rho\Big(\frac{\partial K}{\partial x}W - (\frac{\partial}{\partial t}\frac{\partial L}{\partial \dot{x}})W - ((\frac{\partial}{\partial x}(\frac{\partial L}{\partial \dot{x}}W))V + \frac{\partial L}{\partial \dot{x}}[V,W]\Big)dxdt \\
&= \iint \rho\Big(\frac{\partial K}{\partial x}W - (\frac{\partial}{\partial t}\frac{\partial L}{\partial \dot{x}})W - ((\frac{\partial}{\partial x}\frac{\partial L}{\partial \dot{x}})W)V - (\frac{\partial L}{\partial \dot{x}}\frac{\partial W}{\partial x})V \\
&+ \frac{\partial L}{\partial \dot{x}}(\frac{\partial W}{\partial x}V - \frac{\partial V}{\partial x}W)\Big)dxdt \\
&= \iint \rho\Big(\frac{\partial K}{\partial x} - \frac{\partial}{\partial t}\frac{\partial L}{\partial \dot{x}} - (\frac{\partial}{\partial x}\frac{\partial L}{\partial \dot{x}})V - \frac{\partial L}{\partial \dot{x}}\frac{\partial V}{\partial x}\Big)Wdxdt.
\end{aligned}$$

But $\frac{\partial K}{\partial x} = \frac{\partial L}{\partial x} + \frac{\partial L}{\partial \dot{x}}\frac{\partial V}{\partial x}$, and thus

$$\delta S_1 = \frac{dS_1}{ds}|_{s=0} = \iint \rho\Big(\frac{\partial L}{\partial x} - (\frac{\partial}{\partial t} + V\frac{\partial}{\partial x})\frac{\partial L}{\partial \dot{x}}\Big)Wdxdt.$$

By applying integration by parts and the fact that ρ and thus its derivatives vanish at infinity, we have

$$\begin{aligned}
\frac{dS_2}{ds} &= \iint \Big(\frac{\partial \rho}{\partial s}F + \rho(\frac{\partial F}{\partial y}\frac{\partial \rho}{\partial s} + \frac{\partial F}{\partial y_i}\frac{\partial^2 \rho}{\partial s \partial x_i} + \frac{\partial F}{\partial y_{ij}}\frac{\partial^3 \rho}{\partial s \partial x_i \partial x_j} + \cdots)\Big)dxdt \\
&= \iint \Big(\frac{\partial \rho}{\partial s}F + \rho\frac{\partial F}{\partial y}\frac{\partial \rho}{\partial s} - (\frac{\partial}{\partial x_i}(\rho\frac{\partial F}{\partial y_i}))\frac{\partial \rho}{\partial s} + (\frac{\partial^2}{\partial x_i \partial x_j}(\rho\frac{\partial F}{\partial y_{ij}}))\frac{\partial \rho}{\partial s} - \cdots\Big)dxdt \\
&= \iint \frac{\partial \rho}{\partial s}\Big(F + \rho\frac{\partial F}{\partial y} - \frac{\partial}{\partial x_i}(\rho\frac{\partial F}{\partial y_i}) + \frac{\partial^2}{\partial x_i \partial x_j}(\rho\frac{\partial F}{\partial y_{ij}}) - \cdots\Big)dxdt \\
&= \iint \rho\frac{\partial}{\partial x}\Big(F + \rho\frac{\partial F}{\partial y} - \frac{\partial}{\partial x_i}(\rho\frac{\partial F}{\partial y_i}) + \frac{\partial^2}{\partial x_i \partial x_j}(\rho\frac{\partial F}{\partial y_{ij}}) - \cdots\Big)Wdxdt.
\end{aligned}$$

Thus

$$\frac{dS}{ds}|_{s=0} = \frac{dS_1}{ds}|_{s=0} + \frac{dS_2}{ds}|_{s=0},$$

for all W implies the desired result. □

The following fact is due to Iranian theoretical physicists Mehdi Golshani, Mahdi Atiq and Mozafar Karamian, [13],[14] which we have adapted with our theory.

Corollary 3.3. *When* $L(x,v) = \frac{1}{2m}\|v\|^2 - U(x)$ *and*

$$F(y, y_i, y_{ij}) = -\frac{\hbar^2}{2m}(-\frac{\sum_i y_i^2}{4y^2} + \frac{\sum_i y_{ii}}{2y}),$$

then first of all the identity (3.4) holds for all ρ *and secondly if in the weak Euler-Lagrange equation the vector field* V *is of the form* $V = \frac{\nabla S}{m}$ *for some function* S *then the weak Euler-Lagrange equations convert to the Schrödinger equation (1.1), under the transformation*

$$\psi := \sqrt{\rho}e^{\frac{i}{\hbar}S}. \tag{3.6}$$

In this case we have,

$$F(\rho, \partial_i\rho, \partial^2_{ij}\rho) = -\frac{\hbar^2}{2m}\frac{\nabla^2\sqrt{\rho}}{\sqrt{\rho}}, \tag{3.7}$$

which is called quantum potential in the literature of Bohmian quantum mechanics.

Proof: We have $\frac{\partial F}{\partial y} = \frac{1}{2}y^{-3}\sum_i y_i^2 - \frac{1}{2}y^{-2}\sum_i y_{ii}$, $\frac{\partial F}{\partial y_i} = \frac{-1}{2}y^{-2}y_i$ and $\frac{\partial F}{\partial y_{ij}} = \frac{1}{2}\delta_{ij}y^{-1}$. Thus $\frac{\partial F}{\partial y}(\rho, \partial_i\rho, \partial^2_{ij}\rho) = \frac{1}{2}\rho^{-3}\sum_i(\partial_i\rho)^2 - \frac{1}{2}\rho^{-1}\sum_i \partial^2_{ii}\rho$, $\frac{\partial F}{\partial y_i}(\rho, \partial_i\rho, \partial^2_{ij}\rho) = \frac{-1}{2}\rho^{-2}\partial_i\rho$ and $\frac{\partial F}{\partial y_{ij}}(\rho, \partial_i\rho, \partial^2_{ij}\rho) = \frac{1}{2}\delta_{ij}\rho^{-1}$. Hence,

$$\begin{aligned}
&\rho\frac{\partial F}{\partial y}(\rho, \partial_i\rho, \partial^2_{ij}\rho) - \sum_i \partial_i(\rho\frac{\partial F}{\partial y_i}(\rho, \partial_i\rho, \partial^2_{ij}\rho)) + \sum_{ij}\partial^2_{ij}(\rho\frac{\partial F}{\partial y_{ij}}(\rho, \partial_i\rho, \partial^2_{ij}\rho))\\
=\ & \frac{1}{2}\rho^{-2}\sum_i(\partial_i\rho)^2 - \frac{1}{2}\rho^{-1}\sum_i\partial^2_{ii}\rho + \frac{1}{2}\sum_i(\partial_i\rho^{-1}\partial_i\rho)\\
=\ & \frac{1}{2}\rho^{-2}\sum_i(\partial_i\rho)^2 - \frac{1}{2}\rho^{-1}\sum_i\partial^2_{ii}\rho + \frac{1}{2}(-\rho^{-2}\sum_i(\partial_i\rho)^2 + \rho^{-1}\sum_i\partial^2_{ii}\rho)\\
=\ & 0.
\end{aligned}$$

Also, one can easily check (3.7).

Next, equation (3.3) in this case becomes

$$\rho\Big((\frac{\partial}{\partial t} + V.\nabla)V + \nabla U + \frac{\hbar^2}{2m}\nabla\frac{\nabla^2\sqrt{\rho}}{\sqrt{\rho}}\Big) = 0.$$

One can easily check that the later together with the equation $\frac{\partial\rho}{\partial t} + \nabla.(\rho V) = 0$ when $V = \frac{\nabla S}{m}$ is equivalent with Schrödinger equation under the transformation (3.6). □

Acknowledgment

I would like to express my very great appreciation to professor Mehdi Golshani, for valuable discussions during my visit to the department of physics at IPM Tehran Iran 2010.

This research was in part supported by grants from IPM Schools of Mathematics (No.1403570115) and Physics (IPM/P-2026/02), Iran.

References

[1] Evans L. C., Partial Differential Equations. Springer, (2010).

[2] Holland P. R., The Quantum Theory of Motion. Cambridge University Press, (1993).

[3] Ambrosio L., Gigli N., Savare G., Gradient Flows in Metric Spaces and in the Space of Probability Measures. Birkhauser Verlag, (2005).

[4] Lott J., Villani C., Ricci curvature for metric-measure spaces via optimal transport. Annals of Mathematics, (2009) Vol. 169, Iss. 3, P. 903–991.

[5] Villani C., Optimal Transport, Old and New. Springer, (2009).

[6] Ambrosio L., Gigli N., Construction of the parallel transport in the Wasserstein space. Methods and Applications of Analysis, (2008) Vol. 15, Iss. 1, P. 1–30.

[7] Gangbo W., Nguyen T., Tudorascu A., Hamilton-Jacobi Equations in the Wasserstein Space. Methods and Applications of Analysis, (2008) Vol. 15, Iss. 2, P. 155–184.

[8] Lott J., Some Geometric Calculations on Wasserstein Space. Communications in Mathematical Physics, (2008) Vol. 277, Iss. 2, P. 423–437.

[9] Lott J., An Intrinsic Parallel Transport in Wasserstein Space. Proceedings of the American Mathematical Society, (2017) Vol. 145, Iss. 12, P. 5329–5340.

[10] Otto F., The geometry of dissipative evolution equations: the porous medium equation. Communications in Partial Differential Equations, (2001) Vol. 26, Iss. 1, P. 101–174.

[11] Gangbo W., Swiech A., Metric viscosity solutions of Hamilton-Jacobi equations depending on local slopes. Calculus of Variations and Partial Differential Equations, (2015) Vol. 54, Iss. 1, P. 1183–1218.

[12] Ambrosio L., Gangbo W., Hamiltonian ODEs in the Wasserstein space of probability measures. Communications on Pure and Applied Mathematics, (2008) Vol. 61, Iss. 1, P. 18–53.

[13] Atiq M., Karamian M., Golshani M., A Quasi-Newtonian Approach to Bohmian Mechanics I: Quantum Potential. Annales de la Fondation Louis de Broglie, (2009) Vol. 34, Iss. 1, P. 67–81.

[14] Atiq M., Karamian M., Golshani M., A New Way for the Extension of Quantum Theory: Non-Bohmian Quantum Potentials. Foundations of Physics, (2009) Vol. 39, Iss. 1, P. 33–44.

Akrami S.E.,
Department of Mathematics, Semnan University,
Semnan, Iran,
School of Mathematics, School of Physics,
Institute for Research in Fundamental Sciences (IPM),
Tehran, Iran
e-mail: akramisa@ipm.ir, akramisa@semnan.ac.ir

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 24–32
DOI: 10.29229/uzmj.2026-2-3

On some existence theorems for quasilinear elliptic equations

Amanova N.

Abstract. Boundary value problems are considered for quasilinear elliptic equations with a principal quasilinear elliptic operator of the second order in the Sobolev space $W_p^2(\Omega)$. The main content of the work is to establish the maximum admissible growth of the subordinate nonlinear operator with respect to the corresponding lower-order derivatives. We consider a class of equations for which an a priori estimate of solutions in the norm of the space $\|u\|_{L_l(\Omega)}$ $(1 \le l < \infty)$ implies an a priori estimate of the solution in the norm of the space $W_p^2(\Omega)$. Based on the theorem on a priori estimates, a general solvability theorem for boundary value problems for quasi-linear elliptic equations is obtained under the condition of the existence of an intermediate a priori estimate $\|u\|_{L_l(\Omega)}$ for the solutions of the corresponding family of boundary value problems.

Keywords: interpolation method, Sobolev space, quasilinear equation
MSC (2020): 53C12, 57R25, 57R35

1. Introduction

Let $R_+ \equiv \{t_0 \in R \,|\, t_0 \ge 0\}$, $x = (x_1, \ldots, x_N)$ be a point in the space R^N, $\Omega \subset R^N$ be a bounded domain with the boundary $\partial\Omega$ of class C^2.

Throughout the paper, the functions are assumed to be real-valued. In the paper, we use the following function spaces (see [1, p. 107], [2] and [3]): the spaces of summable functions $L_p(\Omega)$, $p \ge 1$ with the norm

$$\|u\|_{p:\Omega} = \left(\int\limits_{\Omega} |u(x)|^p \, dx \right)^{1/p};$$

the anisotropic Sobolev space $W_p^2(\Omega)$ with the norm

$$\|u\|_{W_p^2(\Omega)} = \|u\|_{p:\Omega} + \sum_{i=1}^{N} \left\| \frac{\partial^2 u}{\partial x_i^2} \right\|_{p:\Omega}.$$

By $D^i u$ we denote a vector from the partial derivatives $D^\alpha u$ of the function $u(x)$, $\alpha = (\alpha_1, \ldots, \alpha_N)$, $|\alpha| = i$ $(i = 1, 2)$.

Consider the following boundary value problem

$$\begin{cases} Lu = f(x, u, Du), & x \in \Omega, \\ u|_{\partial\Omega} = \varphi(x), & x \in \partial\Omega. \end{cases} \tag{1.1}$$

We consider this problem in the class of real functions from the Sobolev space $W_p^2(\Omega)$ for $p > 1$, so that the boundary function $\varphi(x)$ belongs to the space $W_p^{2-1/p}(\partial\Omega)$ (see [2, p. 389]).

The paper provides a theorem on a priori estimates of the solutions $\|u\|_{W_p^2(\Omega)}$ in the norm of the Sobolev space $W_p^2(\Omega)$, expressed by the norm $\|u\|_{l:\Omega} \le M$ $(1 \le l < \infty)$, with some $M > 0$.

For the boundary value problem (1.1). We assume that the following conditions were fulfilled.

A.1) Let the function $f(x, \xi_0, \xi_1)$ be defined on $\overline{\Omega} \times R \times R^N$ with values in R and satisfy the Caratheodory condition, i.e. let it be measurable with respect to x for all $(\xi_0, \xi_1) \in R \times R^N$ and continuous with respect to (ξ_0, ξ_1) for almost all $x \in \Omega$.

A.2) Let $N > 2$ and

$$|f(x, t, \xi_0, \xi_1)| \le b(x) + \sum_{i=0}^{1} b_i(x) \cdot |\xi_i|^{\mu_i}$$

almost for all $x \in \Omega$ and for all $\xi_0 \in R$, $\xi_1 \in R^N$, with non-negative functions $b(x)$ and $b_i(x)$ $(i = 0, 1)$, such that the function $b(x) \in L_p(\Omega)$, $p > N$, and $b_i(x) \in L_{q_i}(\Omega)$, $p < q_i \leq \infty$, $i = 0, 1$.

A.3) Let

$$\mu_i = \frac{1}{i + N/l} \cdot \left(2 + \frac{N}{l} - \frac{N}{q_i}\right) \quad (i = 0, 1).$$

A.4) Let Lu be a quasilinear elliptic operator of second order of the form

$$Lu = \sum_{|\alpha|=2} a_\alpha(x, u, Du) D^\alpha u,$$

where the coefficients a_α $(|\alpha| = 2)$ of the operator L are real and continuous functions on $\overline{\Omega} \times R \times R^N$.

Let $L_v u$ linear with respect to $u(x)$ operator equal to

$$L_v u = \sum_{|\alpha|=2} a_\alpha(x, v, Dv) D^\alpha u,$$

for any function $v(x) \in C^1(\overline{\Omega})$ be a linear elliptic operator, with such real coefficients that the linear (with respect to $u(x)$) boundary value problem

$$\begin{cases} L_v u = g(x), & x \in \Omega, \\ u|_{\partial\Omega} = \varphi(x), & x \in \partial\Omega \end{cases} \tag{1.2}$$

is coercive in the space $W_p^2(\Omega)$, i.e. a priori estimate

$$\|u\|_{W_p^2(\Omega)} \leq C\left(\|g\|_{p;\Omega} + \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)} + \|u\|_{p;\Omega}\right),$$

with positive constant C, independent of $g \in L_p(\Omega)$, $\varphi(x) \in W_p^{2-1/p}(\partial\Omega)$ and on the solution $u \in W_p^2(\Omega)$ of linear problem (1.2), is fulfilled. This time, the constant C can depend on the modulus of continuity of the coefficients a_α $(|\alpha| = 2)$ and modules of continuity of functions v, Dv (see [4]).

For elliptic operators, this problem was considered in [5]-[6], [7]-[8]. In [5, 6, 7, 2], it is assumed that a nonlinear function $f(x, u, Du)$ is continuous in all its arguments. In this case, the well-known S. N. Bernstein growth condition a non-linear function is sufficient for the a priori estimate in $\|u\|_{\infty:\Omega}$, to follow from the a priori estimate in $\|Du\|_{\infty:\Omega}$ and, hence the estimate in $\|u\|_{W_p^2(\Omega)}$. In [9, 8], a boundary value problem for elliptic equations, respectively, was considered, where the non-linear function f belongs to the space $L_p(\Omega)$. In the present paper, for the non-linear function $f(x, u, Du)$, we adopt a different S. N. Bernstein separation [6] and the first an a priori estimate of $\|u\|_{l:\Omega} \leq M$, $1 \leq l < \infty$ and investigate the problem (1.1) using the interpolation method [9, 8]. In this context, we also refer to the work [10, 11, 12, 13].

To study problem (1.1), we will make essential use of the embedding theorem for spaces $W_p^2(\Omega)$ and the multiplicative inequality that follows from it. Let us formulate the embedding theorem [1, p.278] in the form we need.

Embedding theorems: Assume that $S_i \geq p$ $(i = 0, 1)$ and

$$\varkappa_1 = 1 - \frac{1}{2}\left[N\left(\frac{1}{p} - \frac{1}{S_i}\right) + i\right] > 0, \quad i = 0, 1.$$

Then the space $W_p^2(\Omega) \subset L_{S_i}(\Omega)$ $(i = 0, 1)$ is compactly embedded, and

$$\|D^i u\|_{S_i:\Omega} \leq C \cdot \|u\|_{W_p^2(\Omega)} \quad (i = 0, 1).$$

If $S_i > l$ $(i = 0, 1)$, then there exists $\delta_0 > 0$ such that for any $0 < \delta \leq \delta_0$ and for function $u \in W_p^2(\Omega)$

$$\|D^i u\|_{S_i:\Omega} \leq C \cdot \delta^{-\varkappa_0} \cdot \|u\|_{l;\Omega} + C \cdot \delta^{\varkappa_1} \cdot \|u\|_{W_p^2(\Omega)}, \quad (i = 0, 1), \tag{1.3}$$

where

$$\varkappa_0 = \frac{1}{2}\left[N\left(\frac{1}{p} - \frac{1}{S_i}\right) + i\right], \quad i = 0, 1.$$

Here, the positive constant C does not depend on the function $u(x)$ and δ.
Assume (see [1, p. 392])

$$\delta = \min\left\{\left(\|u\|_{l:\Omega}\cdot\|u\|^{-1}_{W_p^2(\Omega)}\right)^{\frac{1}{\varkappa_0+\varkappa_1}};\delta_0\right\}.$$

Then of simple transformations we have

$$\|D^i u\|_{S_i:\Omega}\le C_0\cdot\|u\|^{1/\mu_i}_{W_p^2(\Omega)}\cdot\|u\|^{1-1/\mu_i}_{l:\Omega}+C_0\cdot\|u\|_{l:\Omega},\quad i=0,1. \tag{1.4}$$

Where $S_i=\frac{\mu_i p q_i}{q_i-p}$, $\frac{\varkappa_0}{\varkappa_0+\varkappa_1}=\frac{1}{\mu_i}$, $i=0,1$ and positive constants C_0 independent of function $u(x)$. Note that under the condition on $S_i>l$ $(i=0,1)$ the space $W_p^2(\Omega)\subset L_{S_i}(\Omega)\subset L_l(\Omega)$ $(i=0,1)$, i.e. $\|u\|_{l:\Omega}\le \text{const}\cdot\|u\|_{W_p^2(\Omega)}$.
These results can be formulated as follows.

Lemma 1.1. *Let the conditions on $S_i\ge p$, $S_i>l$, $i=0,1$ and $\varkappa_1>0$ be fulfilled. Then for an arbitrary function $u(x)\in W_p^2(\Omega)$ the inequality* (1.4) *holds.*

Assume now the $F(u)(x)\equiv f(x,u(x),Du(x))$. Let $u(x)\in W_p^2(\Omega)$, the conditions A.1)–A.3), $\mu_i>\frac{l(q_i-p)}{q_i p}$ $(i=0,1)$ and $\varkappa_1>0$ be fulfilled. From condition A.2) it follows that

$$\|F(u)\|_{p:\Omega}\le\|b\|_{p:\Omega}+\sum_{i=0}^{1}\|b_i\|_{q_i:\Omega}\cdot\|D^i u\|^{\mu_i}_{S_i:\Omega},$$

where $S_i=\frac{\mu_i p q_i}{q_i-p}$ $(i=0,1)$. We will consider the case $S_i>l$. From this we also obtain that $\mu_i>\frac{l(q_i-p)}{q_i p}(i=0,1)$. Then, based on the equalities A.3) and interpolation inequalities (1.4). , we obtain

$$\|F(u)\|_{p:\Omega}\le\|b\|_{p:\Omega}+\sum_{i=0}^{1}\|b_i\|_{q_i:\Omega}\cdot\left[C_0\cdot\|u\|^{1/\mu_i}_{W_p^2(\Omega)}\cdot\|u\|^{1-1/\mu_i}_{l:\Omega}+C_0\cdot\|u\|_{l:\Omega}\right]^{\mu_i}\le$$

$$\le\|b\|_{p:\Omega}+\sum_{i=0}^{1}\|b_i\|_{q_i:\Omega}\cdot 2^{\mu_i-1}\cdot C_0^{\mu_i}\cdot\|u\|^{\mu_i-1}_{l:\Omega}\cdot\|u\|_{W_p^2(\Omega)}+$$

$$+\sum_{i=0}^{1}\|b_i\|_{q_i:\Omega}\cdot 2^{\mu_i-1}\cdot C_0^{\mu_i}\cdot\|u\|^{\mu_i}_{l:\Omega}=$$

$$=\Phi_1(\|u\|_{l:\Omega})\cdot\|u\|_{W_p^2(\Omega)}+\Phi_2(\|u\|_{l:\Omega}). \tag{1.5}$$

Where

$$\Phi_1(\|u\|_{l:\Omega})=\sum_{i=0}^{1}\|b_i\|_{q_i:\Omega}\cdot 2^{\mu_i-1}\cdot C_0^{\mu_i}\cdot\|u\|^{\mu_i-1}_{l:\Omega},$$

$$\Phi_2(\|u\|_{l:\Omega})=\|b\|_{p:\Omega}+\sum_{i=0}^{1}\|b_i\|_{q_i:\Omega}\cdot 2^{\mu_i-1}\cdot C_0^{\mu_i}\cdot\|u\|^{\mu_i}_{l:\Omega},$$

i.e. $\Phi_1,\Phi_2: R_+\to R_+$ are increasing functions determined by the known data. This proves the boundedness of the operator $F(u)$.

On the other hand, $p>N$. Then, by virtue of the Sobolev embedding theorem [14], it follows that the embedding operator $W_p^2(\Omega)\to L_{S_i}(\Omega)$ is completely continuous. By virtue of (1.5), the operator $L_{S_i}(\Omega)\to L_p(\Omega)$ $(i=0,1)$ is bounded and, by the general property of the superposition operator, is continuous. Thus, the operator $W_p^2(\Omega)\to L_p(\Omega)$ is completely continuous as a composition of completely continuous and continuous operators .

Thus, the following lemma is proved.

Lemma 1.2. *Let $u(x)\in W_p^2(\Omega)$ and the conditions A.1) - A.3), $\mu_i>\frac{l(q_i-p)}{pq_i}$, $(i=0,1)$, $\varkappa_1>0$ be fulfilled. Then the operator $F(u)(x)$ is a bounded and continuous operator from $W_p^2(\Omega)$ to $L_p(\Omega)$.*

Theorem 1.3. *Let conditions A.1) - A.3) be fulfilled and $p > N$. Then there exists a function $\Phi : R_+^2 \to R_+$ increasing in each argument such for any possible solution $u \in W_p^2(\Omega)$ of problem (1.1) the a priori estimate*

$$\|u\|_{W_p^2(\Omega)} \leq \Phi\left(M, \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)}\right)$$

holds. Where, $\|u\|_{l;\Omega} \leq M$, $l \in [1, +\infty)$ and $R_+^2 = R_+ \times R_+$.

Proof: It is evident that

$$Lu = f(x, u, Du) = \frac{f(x, u, Du)}{1 + |u|^{\mu_0} + |Du|^{\mu_1}} \cdot \left(1 + |u|^{\mu_0} + |Du|^{\mu_1}\right).$$

Hence, for the function $u(x)$

$$\begin{cases} Lu - C(x)\, u = f_1(x) \cdot |u|^{\mu_0} + f_1(x) \cdot |Du|^{\mu_1} + f_0(x), & x \in \Omega, \\ u\Big|_{\partial\Omega} = \varphi(x), & x \in \partial\Omega. \end{cases} \tag{1.6}$$

Where, $C(x) = b(x) \geq 0$, $C(x) \in L_p(\Omega)$,

$$f_1(x) = \frac{f(x, u, Du)}{1 + |u|^{\mu_0} + |Du|^{\mu_1}}, \quad f_0(x) = f_1(x) - C(x)\, u(x),$$

$f_0(x) \in L_p(\Omega)$, $S_0 = \frac{\mu_0 p q_0}{q_0 - p}$ when $f_1(x) \in L_{q_0}(\Omega)$, $S_1 = \frac{\mu_1 p q_1}{q_1 - p}$ when $f_1(x) \in L_{q_1}(\Omega)$; $q_0 > p > N$, $q_1 > p > N$ and $\mu_i = \frac{1}{i + N/l} \cdot \left(2 + \frac{N}{l} - \frac{N}{q_i}\right)$ $(i = 0, 1)$.

Consider in the space $W_p^2(\Omega)$ with $p > N$ the boundary value problem.

$$\begin{cases} LV - C(x)\, V(x) = f_1(x)\left(|V|^{\mu_0} + |Du|^{\mu_1}\right) + \lambda \cdot f_0(x), & x \in \Omega, \\ V(x)\Big|_{\partial\Omega} = \lambda \cdot \varphi(x), & x \in \partial\Omega, \end{cases} \tag{1.7}$$

with the problem $\lambda \in [0, 1]$ and the functions $C(x), \varphi(x)$, $f_0(x)$ and $f_1(x)$ defined above. It should be noted that a problem of the form (1.7) has also been considered in [5, 10, 9]. □

Lemma 1.4. *For any fixed $\lambda \in [0, 1]$, the problem (1.7) has at most one solution $u(x)$ from $W_p^2(\Omega)$, $p > N$.*

Proof: Assume the contrary. Then for the difference $\mathcal{W}(x) = V(x) - Z(x)$ of two possible solutions $V(x)$ and $Z(x)$,,we have

$$\begin{cases} L\mathcal{W}(x) - C(x)\, \mathcal{W}(x) = f_1(x)\left[\sum\limits_{i=1}^{N} h_i(x)\, D_i\mathcal{W}(x) + h(x)\, \mathcal{W}(x)\right], & x \in \Omega, \\ \mathcal{W}(x)\Big|_{\partial\Omega} = 0, & x \in \partial\Omega. \end{cases}$$

Here,

$$h_i(x) = \int_0^1 H_i(x, \tau)\, d\tau, \quad h(x) = \int_0^1 H(x, \tau)\, d\tau.$$

$$H_i(x, \tau) = \mu_1 \left[\sum_{k=1}^{N} \left(\tau \cdot D_k\mathcal{W} + D_k Z\right)^2\right]^{\frac{\mu_1}{2} - 1} \left(\tau \cdot D_i\mathcal{W} + D_i Z\right)(x).$$

for

$$\sum_{k=1}^{N} \left(\tau \cdot D_k\mathcal{W} + D_k Z\right)^2 (x) \neq 0;$$

$$H(x, \tau) = \mu_0 \left[\left(\tau\mathcal{W} + Z\right)^2\right]^{\frac{\mu_0}{2} - 1} \left(\tau\mathcal{W} + Z\right)(x) \quad \text{for} \quad \left(\tau\mathcal{W} + Z\right)^2 \neq 0$$

and

$$H_i(x,\tau)=0 \quad \text{for} \quad \sum_{k=1}^{N}\left(\tau\cdot D_k\mathcal{W}+D_kZ\right)^2(x)=0,$$

$$H(x,\tau)=0 \quad \text{for} \quad \left(\tau\cdot\mathcal{W}+Z\right)^2(x)=0.$$

Since $C(x)\geq 0$, $C(x)\in L_p(\Omega)$,$h(x)\geq 0$, $h(x)\cdot f_1(x)\in L_p(\Omega)$, $f_1(x)\cdot h_i(x)\in L_{q_1}(\Omega)$ $(i=1,\ldots,N)$, and $p>N$, from the results of [15] it follows that $\mathcal{W}\equiv 0$ in Ω.

Lemma 1.4 is proved. □

Now let $v_1(x)$ and $v_2(x)$ be solutions to problem (1.7) corresponding to the values λ_1 and λ_2 $(\lambda_2>\lambda_1)$ of the parameter λ respectively. Then for the difference $\tilde{V}(x)=v_2(x)-v_1(x)$, we have

$$\begin{cases} L\tilde{V}-C(x)\cdot\tilde{V}(x)=f_1(x)\left[\sum\limits_{i=1}^{N}\tilde{h}_i(x)D_i\tilde{V}+\tilde{h}(x)\tilde{V}\right]+\\ +(\lambda_2-\lambda_1)f_0(x), \quad x\in\Omega,\\ \tilde{V}\Big|_{\partial\Omega}=(\lambda_2-\lambda_1)\cdot\varphi(x), \quad x\in\partial\Omega, \end{cases}$$

where

$$\tilde{h}_i(x)=\int\limits_0^1\tilde{H}_i(x,\tau)\,d\tau, \quad \tilde{h}(x)=\int\limits_0^1\tilde{H}(x,\tau)\,d\tau.$$

$$\tilde{H}_i(x,\tau)=\mu_1\left[\sum_{k=1}^{N}\left(\tau\cdot D_k\tilde{V}+D_kv_1\right)^2\right]^{\frac{\mu_1}{2}-1}\cdot\left(\tau\cdot D_i\tilde{V}+D_iv_1\right)(x),$$

for

$$\sum_{k=1}^{N}\left(\tau\cdot D_k\tilde{V}+D_kv_1\right)^2(x)\neq 0, \quad \tilde{H}_i(x,\tau)=0$$

for

$$\sum_{k=1}^{N}\left(\tau\cdot D_k\tilde{V}+D_kv_1\right)^2(x)=0;$$

$$\tilde{H}(x,\tau)=\mu_0\left[\left(\tau\tilde{V}+v_1\right)^2\right]^{\frac{\mu_0}{2}-1}\cdot\left(\tau\tilde{V}+v_1\right)(x).$$

for $\left(\tau\cdot\tilde{V}+v_1\right)^2(x)\neq 0$, $\tilde{H}(x,\tau)=0$ for $\left(\tau\cdot\tilde{V}+v_1\right)^2(x)=0$.

Assume

$$\mathbb{K}=(\lambda_2-\lambda_1)\left(1+\|u\|_{l:\Omega}\right), \quad \lambda_2>\lambda_1.$$

Lemma 1.5. $\left\|\tilde{V}(x)\right\|_{l:\Omega}\leq\mathbb{K}$.

Proof: For the function $\left(\tilde{V}-\mathbb{K}\right)$, we have

$$\begin{cases} L\left(\tilde{V}-\mathbb{K}\right)-C(x)\left(\tilde{V}-\mathbb{K}\right)=f_1(x)\left[\sum\limits_{i=1}^{N}\tilde{h}_i(x)D_i\left(\tilde{V}-\mathbb{K}\right)+\tilde{h}(x)\left(\tilde{V}-\mathbb{K}\right)\right]+\\ +(\lambda_2-\lambda_1)f_0(x)+C(x)\mathbb{K}+f_1(x)\tilde{h}(x)\mathbb{K}, \quad x\in\Omega,\\ \left(\tilde{V}-\mathbb{K}\right)\Big|_{\partial\Omega}=(\lambda_2-\lambda_1)\cdot\varphi(x)-\mathbb{K}, \quad x\in\partial\Omega. \end{cases}$$

Further,

$$(\lambda_2-\lambda_1)f_0(x)+C(x)\mathbb{K}+f_1(x)\tilde{h}(x)\mathbb{K}=$$

$$=(\lambda_2-\lambda_1)\cdot\left[f_0(x)+\left(C(x)+f_1(x)\tilde{h}(x)\right)\left(1+\|u\|_{l:\Omega}\right)\right]=$$

$$= (\lambda_2 - \lambda_1) \cdot \left[f_1(x) - C(x) u(x) + \left(C(x) + f_1(x) \tilde{h}(x) \right) (1 + \|u\|_{l:\Omega}) \right] =$$
$$= (\lambda_2 - \lambda_1) \cdot [C(x) (\|u\|_{l:\Omega} - u(x)) + f_1(x) +$$
$$+ C(x) + f_1(x) \tilde{h}(x) (1 + \|u\|_{l:\Omega}) \Big] \geq 0, \quad x \in \Omega;$$
$$\left(\tilde{V} - \mathbb{K} \right) \Big|_{\partial\Omega} = (\lambda_2 - \lambda_1) \varphi(x) - (\lambda_2 - \lambda_1) (1 + \|u\|_{l:\Omega}) \leq$$
$$\leq (\lambda_2 - \lambda_1) (\varphi(x) - \|u\|_{l:\Omega}) \leq 0, \quad x \in \partial\Omega.$$

Then from [15] it follows that $\tilde{V}(x) \leq \mathbb{K}$ in $\overline{\Omega}$. Similarly, inequality $\tilde{V}(x) \geq -\mathbb{K}$ is proved in $\overline{\Omega}$. Lemma 1.5 is proved. □

Consider a parametric family of problems (1.7). It should be noted that, by Lemma 1.4, the solution of problem (1.6) and (in view of the adopted notation) with the solution of problem (1.1).

Let $v_1(x)$ and $v_2(x)$ be solutions to problem (1.7), corresponding to the values λ_1 and λ_2 ($\lambda_2 > \lambda_1$) of the parameter λ. Then, by Lemma 1.5 we have

$$\|v_2 - v_1\|_{l:\Omega} \leq (\lambda_2 - \lambda_1) (1 + \|u\|_{l:\Omega}). \tag{1.8}$$

On the other hand, this function is a solution to the problem

$$\begin{cases} L\tilde{V} - C(x) \tilde{V} = f_1(x) [(|v_2|^{\mu_0} - |v_1|^{\mu_0}) + (|Dv_2|^{\mu_1} - |Dv_1|^{\mu_1})] + \\ + (\lambda_2 - \lambda_1) f_0(x), \quad x \in \Omega, \\ \tilde{V}\Big|_{\partial\Omega} = (\lambda_2 - \lambda_1) \varphi(x), \quad x \in \partial\Omega. \end{cases}$$

Consequently,

$$\left\| L\tilde{V} - C(x) \tilde{V} \right\|_{p:\Omega} \leq 2^{\mu_0 - 1} \cdot \|f_1\|_{q_0:\Omega} \cdot \left\| \tilde{V} \right\|^{\mu_0}_{S_0:\Omega} +$$
$$+ 2^{\mu_1 - 1} \cdot \|f_1\|_{q_1:\Omega} \cdot \left\| D\tilde{V} \right\|^{\mu_1}_{S_1:\Omega} + 2^{\mu_0} \cdot \|f_1\|_{q_0:\Omega} \cdot \|v_1\|^{\mu_0}_{S_0:\Omega} +$$
$$+ 2^{\mu_1} \cdot \|f_1\|_{q_1:\Omega} \cdot \|Dv_1\|^{\mu_1}_{S_1:\Omega} + (\lambda_2 - \lambda_1) \|f_0\|_{p:\Omega}, \tag{1.9}$$

$$\left\| \tilde{V} \right\|_{W_p^{2-1/p}(\partial\Omega)} \leq (\lambda_2 - \lambda_1) \cdot \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)}. \tag{1.10}$$

On the other hand, it follows from (1.4) that

$$\left\| D^i \tilde{V} \right\|_{S_i:\Omega} \leq C_0 \cdot \left\| \tilde{V} \right\|^{1/\mu_i}_{W_p^2(\Omega)} \cdot \left\| \tilde{V} \right\|^{1 - 1/\mu_i}_{l:\Omega} + C_0 \cdot \left\| \tilde{V} \right\|_{l:\Omega}, \quad i = 0, 1. \tag{1.11}$$

By virtue of the well-known linear theory of elliptic problems, the inequality

$$\left\| \tilde{V} \right\|_{W_p^2(\Omega)} \leq C_2 \left(\left\| L\tilde{V} - C(x) \tilde{V} \right\|_{p:\Omega} + \left\| \tilde{V} \right\|_{W_p^{2-1/p}(\partial\Omega)} \right).$$

Here, $C_2 = C_2(\Omega, p, q_0, q_1, N)$ is a positive constant. Then, using inequalities (1.8)–(1.11) we obtain

$$\left\| \tilde{V} \right\|_{W_p^2(\Omega)} \leq C_2 \cdot 2^{\mu_0 - 1} \cdot \|f_1\|_{q_0:\Omega} \cdot \left\| \tilde{V} \right\|^{\mu_0}_{S_0:\Omega} +$$
$$+ C_2 \cdot 2^{\mu_1 - 1} \cdot \|f_1\|_{q_1:\Omega} \cdot \left\| D\tilde{V} \right\|^{\mu_1}_{S_1:\Omega} +$$
$$+ C_2 \left(\|f_1\|_{q_0:\Omega} \cdot 2^{\mu_0} \cdot \|v_1\|^{\mu_0}_{S_0:\Omega} + \|f_1\|_{q_1:\Omega} \cdot 2^{\mu_1} \cdot \|Dv_1\|^{\mu_1}_{S_1:\Omega} \right) +$$
$$+ C_2 (\lambda_2 - \lambda_1) \left(\|f_0\|_{p:\Omega} + \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)} \right) \leq$$

$$\leq C_2 \cdot 2^{\mu_0-1} \cdot \|f_1\|_{q_0:\Omega} \cdot \left[C_0 \cdot \left\|\tilde{V}\right\|^{1/\mu_0}_{W_p^2(\Omega)} \cdot \left\|\tilde{V}\right\|^{1-1/\mu_0}_{l:\Omega} + C_0 \cdot \left\|\tilde{V}\right\|_{l:\Omega}\right]^{\mu_0} +$$

$$+C_2 \cdot 2^{\mu_1-1} \cdot \|f_1\|_{q_1:\Omega} \cdot \left[C_0 \cdot \left\|\tilde{V}\right\|^{1/\mu_1}_{W_p^2(\Omega)} \cdot \left\|\tilde{V}\right\|^{1-1/\mu_1}_{l:\Omega} + C_0 \cdot \left\|\tilde{V}\right\|_{l:\Omega}\right]^{\mu_1} +$$

$$+C_2 \left(\|f_1\|_{q_0:\Omega} \cdot 2^{\mu_0} \cdot \|v_1\|^{\mu_0}_{S_0:\Omega} + 2^{\mu_1} \cdot \|f_1\|_{q_1:\Omega} \cdot \|Dv_1\|^{\mu_1}_{S_1:\Omega}\right) +$$

$$+C_2 (\lambda_2 - \lambda_1) \left(\|f_0\|_{p:\Omega} + \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)}\right) \leq$$

$$\leq C_2 \cdot \left\|\tilde{V}\right\|_{W_p^2(\Omega)} \cdot \left[\|f_1\|_{q_0:\Omega} \cdot 2^{2\mu_0-2} \cdot C_0^{\mu_0} \cdot (\lambda_2 - \lambda_1)^{\mu_0-1} \cdot (1 + \|u\|_{l:\Omega})^{\mu_0-1} +\right.$$

$$\left. + \|f_1\|_{q_1:\Omega} \cdot 2^{2\mu_1-2} \cdot C_0^{\mu_1} \cdot (\lambda_2 - \lambda_1)^{\mu_1-1} \cdot (1 + \|u\|_{l:\Omega})^{\mu_1-1}\right] +$$

$$+C_2 \left(\|f_1\|_{q_0:\Omega} \cdot 2^{\mu_0} \cdot \|v_1\|^{\mu_0}_{S_0:\Omega} + \|f_1\|_{q_1:\Omega} \cdot 2^{\mu_1} \cdot \|Dv_1\|^{\mu_1}_{S_1:\Omega}\right) +$$

$$+C_2 \cdot 2^{2\mu_0-2} \cdot \|f_1\|_{q_0:\Omega} \cdot C_0^{\mu_0} \cdot (\lambda_2 - \lambda_1)^{\mu_0} \cdot (1 + \|u\|_{l:\Omega})^{\mu_0} +$$

$$+C_2 \cdot 2^{2\mu_1-2} \cdot \|f_1\|_{q_1:\Omega} \cdot C_0^{\mu_1} \cdot (\lambda_2 - \lambda_1)^{\mu_1} \cdot (1 + \|u\|_{l:\Omega})^{\mu_1} +$$

$$+C_2 (\lambda_2 - \lambda_1) \left(\|f_0\|_{p:\Omega} + \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)}\right).$$

Let

$$C_2 \cdot \left(\|f_1\|_{q_0:\Omega} \cdot 2^{2\mu_0-2} \cdot C_0^{\mu_0} \cdot (\lambda_2 - \lambda_1)^{\mu_0-1} \cdot (1 + \|u\|_{l:\Omega})^{\mu_0-1} +\right.$$

$$\left. + \|f_1\|_{q_1:\Omega} \cdot 2^{2\mu_1-2} \cdot C_0^{\mu_1} \cdot (\lambda_2 - \lambda_1)^{\mu_1-1} \cdot (1 + \|u\|_{l:\Omega})^{\mu_1-1}\right) \leq \frac{1}{2}.$$

Then, if

$$0 < \lambda_2 - \lambda_1 \leq h \tag{1.12}$$

holds, we obtain the inequality

$$\left\|\tilde{V}\right\|_{W_p^2(\Omega)} \leq \Phi_1 + \Phi_2 \cdot \|v_1\|^{\mu_0}_{S_0:\Omega} + \Phi_3 \cdot \|Dv_1\|^{\mu_1}_{S_1:\Omega}. \tag{1.13}$$

Here, the functions $\Phi_1, \Phi_2, \Phi_3 : R_+ \to R_+$ are increasing functions defined by the data of the problem (1.1), the quantities $\lambda_1, \lambda_2 \in [0,1]$ satisfy the condition (1.12), and the functions $v_1(x), v_2(x) \in W_p^2(\Omega)$ are solutions corresponding to the quantities λ_1 and λ_2 of the problem (1.7).

Denote $\lambda^{(k-1)} = \lambda_1$, $\lambda^{(k)} = \lambda_2$, $v^{(k-1)} = v_1$, and $v^{(k)} = v_2$. From the inequality (1.13), we obtain

$$\left\|v^{(k)}\right\|_{W_p^2(\Omega)} \leq \Phi_1 + \Phi_2 \cdot \left\|v^{(k-1)}\right\|^{\mu_0}_{S_0:\Omega} + \Phi_3 \cdot \left\|Dv^{(k-1)}\right\|^{\mu_1}_{S_1:\Omega} + \left\|v^{(k-1)}\right\|_{W_p^2(\Omega)},$$

for $0 < \lambda^{(k)} - \lambda^{(k-1)} \leq h$, $\lambda^{(k-1)}, \lambda^{(k)} \in [0,1]$,. Hence, by virtue of the independence of h of the indicated values $\lambda^{(k-1)}, \lambda^{(k)}$ and the embedding inequality (see [14]) with $p > N$

$$\left\|v^{(k-1)}\right\|_{S_0:\Omega} \leq \text{const} \cdot \left\|v^{(k-1)}\right\|_{W_p^2(\Omega)},$$

$$\left\|Dv^{(k-1)}\right\|_{S_1:\Omega} \leq \text{const} \cdot \left\|v^{(k-1)}\right\|_{W_p^2(\Omega)},$$

after a finite number of iterations, we obtain the assertion of Theorem 1.3/ In this case, the first iteration ($k = 1$) corresponds to $\lambda^{(0)} = 0$ and $v^{(0)} = 0$.

Theorem 1.3 is proved.

2. Solvability theory

Let us consider boundary value problem (1.1) with real Sobolev space $W_p^2(\Omega)$ with $p > N$ provided that there exists a priori estimate $\|u\|_{l,\Omega} \leq M$ independent of $\lambda \in [0,1]$, for solving parametric family of problem

$$\begin{cases} Lu = \lambda \cdot f(x, u, Du), & x \in \Omega, \\ u\Big|_{\partial\Omega} = \varphi(x), & x \in \partial\Omega \end{cases} \tag{2.1}$$

for $l \in [1, \infty)$ and $M > 0$.

For the operator L and boundary function $\varphi(x)$ we suppose that the following condition is fulfilled.

A.5) Let L be a quasilinear elliptic operator with real continuous coefficients, Suppose that the operator

$$Lu = \sum_{|\alpha|=2} a_\alpha(x, u, Du) D^\alpha u$$

is a second-order $\varphi(x) \in W_p^{2-1/p}(\partial\Omega)$ and linear boundary value problem (1.2) for any $g(x) \in L_p(\Omega)$ is uniquely solvable in space $W_p^2(\Omega)$ so that estimation

$$\|u\|_{W_p^2(\Omega)} \leq C\left(\|g\|_{p:\Omega} + \|\varphi\|_{W_p^{2-1/p}(\partial\Omega)}\right)$$

with positive constant C, independent of $g \in L_p(\Omega)$ and of the solution $u(x) \in W_p^2(\Omega)$ is fulfilled.

Theorem 2.1. *Let conditions A.1)-A.3), A.5) be fulfilled, and for parametric family of problems* (2.1) *for* $\lambda \in [0,1]$ *there exist a priori estimate* $\|u\|_{l:\Omega}$ *, independent of* λ *with* $l \in [1, +\infty)$ *and* $M > 0$ *. Then there exists a solution of problem* (1.1) *in the space* $W_p^2(\Omega)$ *with* $p > N$.

Proof. Let us consider a parametric family of problem (2.1) for $\lambda \in [0,1]$. For this family of problems, the conditions of Theorem 1.3 are fulfilled, by virtue of which there exsists such a constant C_1, independent of λ, that for any possible solution $W_p^2(\Omega)$ of problem (2.1) the inequality

$$\|u\|_{W_p^2(\Omega)} \leq C_1, \quad \forall\, \lambda \in [0,1] \tag{2.2}$$

is fulfilled.

From condition A.5) it follows that linear boundary value problem (1.2) is uniquely solvable and the following representation is valid for iss solutions $\tilde{u}(x)$

$$\tilde{u}(x) = V(x) + \Phi(x), \tag{2.3}$$

where $V(x), \Phi(x) \in W_p^2(\Omega)$ are such that

$$\begin{cases} L_v V = g(x), & x \in \Omega, \\ V\Big|_{\partial\Omega} = 0, & x \in \partial\Omega, \end{cases} \tag{2.4}$$

and

$$\begin{cases} L_v \Phi = 0, & x \in \Omega, \\ \Phi\Big|_{\partial\Omega} = \varphi(x), & x \in \partial\Omega. \end{cases}$$

Linear problem (2.4) is uniquely solvably for any function $g(x) \in L_p(\Omega)$, so that $V = Ag$, where A is a linear continuous operator from $L_p(\Omega)$ to $W_p^2(\Omega)$.. Then, the boundary value problem (2.1) is equivalent to the following operator equation

$$\mathcal{W} = \lambda \cdot T(\mathcal{W}), \quad \lambda \in [0,1] \tag{2.5}$$

consider in the space

$$X = \left\{\mathcal{W}(x) \in W_p^2(\Omega) \;\Big|\; \mathcal{W}\Big|_{\partial\Omega} = 0\right\}.$$

Here,

$$T(\mathcal{W}) = A\tilde{f}(V+\Phi), \quad \tilde{f}(u)(x) = f(x, u(x), Du(x))$$

The operator $\tilde{f}$ was determined in the space $W_p^2(\Omega)$ with the values in $L_p(\Omega)$ and is a completely continuous operator from $W_p^2(\Omega)$ to $L_p(\Omega)$. This fact follows from the Sobolev -Kondrashev embedding theorem [14] based on the conditions of Theorem 2.1. Therefore, the operator T is a completely continuous operator acting in the space X. For possible solutions $\mathcal{W}(x)$ from the space X, by virtue of estimation (2.2) and representation (2.3) the following inequality is fulfilled

$$\|\mathcal{W}\|_X \leq C_2, \quad \forall\ \lambda \in [0,1]$$

where, C_2 is a positive constant independent neither of $W(x)$, nor of λ. Then, by virtue of the Leray-Schauder theorem [3, p.135] equations (2.5) for $\lambda = 1$ and consequently boundary value problem (1.1) have the solution $u_0(x) \in W_p^2(\Omega)$.

The proof of Theorem 2.1 is complete.

References

[1] Besov O. A., Ilin V. A., Nikolski S. M., Integral representations of functions and embeding theorems. N., M., (1996).

[2] Ladyzhenskaya O. A., Uraltseva N. N., Linear and quasilinear equations of elliptic type. Nauka, Moscow, (1973).

[3] Bilalov B. T., Sadiqova S. R., On solvability in the small of higher order elliptic equations in rearrangement invariant spaces. Siberian j. math. (2022) Vol. 63, P. 425–437.

[4] Agmon S., Douglis A., Nirenberg L., Estimations of the solutions of elliptic equations near the boundary. N., M., (1962).

[5] Amann H., Grandall M. G., On some existence theoremes for semilinear elliptic equations. Ind. Univ. Math. j. (1978) Vol. 27, Iss. 5, P. 779–790.

[6] Bernstein S. N., On the equations of the calculus of variations. Collerted works. USSR Academy of Sciences, (1960), P. 191–241.

[7] Kazdan A. I., Kramer R. I., Invariant criteria for existence of solutions to second order quasilinear elliptic equations. Comm. Pure Appl. Math. (1978) Vol. 31, Iss. 5, P. 319–645.

[8] Pohozhaev S. I., On the solvability of quasilinear elliptic equations of arbitrary order. Mathem. Sb. (1982) Vol. 117(159), Iss. 2, P. 251–265.

[9] Pohozhaev S. I., On the equations of the form $\Delta u = f(x, u, Du)$. Matem. Sb. (1980) Vol. 113 (155), Iss. 2 (10), P. 324–338.

[10] Amanov R. A., Ismailov A. I., On equations of the form $\Delta u - \frac{\partial u}{\partial t} = f(x, t, u, Du)$. Azerbaijan Journal of Mathematics, (2023) Vol. 17, Iss. 2, P. 137–158.

[11] Amanova N. R., On quasilinear parabolic equations of higher order. Azerbaijan Journal of Mathematics, (2025) Vol. 15, Iss. 1, P. 191–209.

[12] Amanov R. A., Amanova N. R., On solvability of some quasilinear parabolic equations of higher order. Azerbaijan Journal of Mathematics, (2025) Vol. 15, Iss. 2, P. 240–263.

[13] Khalilov V. S., Amanova N. R., On the cristence of strong solutions of higher order quasilinear. Uzbek Mathematical journal, (2024) Vol. 68, Iss. 1, P. 11–18.

[14] Sobolev S. L., Some applications of functional analysis in mathematical physics. SO AN SSSR, Novosibirsk, (1962).

[15] Aleksandrov A. D., Uniqueness conditions and estimates of the solution to the Dirichlet problem. Vestnik LGU, ser. math. mech. and astr. (1963) Vol. 3, Iss. 13, P. 5–29.

Amanova N.R.,
Baku State University,
Baku, Azerbaijan
e-mail: amanovan93@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 33–41
DOI: 10.29229/uzmj.2026-2-4

Solvability of a nonlinear elliptic problem involving $p-$Laplacian with Dirichlet condition

Benaissa S., Messelmi F., Merouani A.

Abstract. In this work, we study a class of nonlinear elliptic equations of the form $Au = -\Delta_p u + div\,(Vu) = R(x,u)$ in a bounded domain $\Omega \subset \mathbb{R}^n$, where $-\Delta_p u = -div\left(|\nabla u|^{p-2}\nabla u\right)$ denotes the p-Laplacian operator, $V(x)$ is a given vector field, and $R(x)$ represents a source term. We establish the existence and uniqueness of weak solutions under suitable assumptions on V and R.To establish the existence of weak solutions to the nonlinear elliptic problem we employ a fixed point approach based on the compactness of the associated operator.The existence result is obtained by proving that the composition $A^{-1}N_R$ is a compact and continuous mapping that admits at least one fixed point, where N_R is the Nemyskii operator corresponding to a Carathéodory function R . This ensures the existence of a weak solution $u \in W_0^{1,p}(\Omega)$. For the uniqueness of solutions, we rely on the strict monotonicity of the p-Laplacian operator and appropriate growth and Lipschitz conditions on $R(x,u)$. Under these assumptions, one can show that if u_1 and u_2 are two weak solutions, then their difference satisfies an energy inequality leading to $u_1 = u_2$ almost everywhere in Ω.

Keywords: $p-$Laplacian, Fixed point, Variational formulation, Carathéodory function, Nemytskii operator.
MSC (2020): 35J62, 35J92, 47J35

1. Introduction

Partial differential equations (PDEs) play a fundamental role in modeling a wide range of physical, biological, and engineering phenomena. They provide an essential mathematical framework for describing diffusion, transport, and nonlinear interactions that arise in various natural and applied processes. Among them, nonlinear elliptic equations [1, 2], particularly those involving the p-Laplacian operator, have attracted considerable attention due to their rich mathematical structure and broad spectrum of applications in physics, mechanics, and materials science Albo [3, 4, 5, 6, 7, 8, 9, 10].

It is more investigated problem with $divVu = 0$ [11, 4, 5, 6, 7], and the references therein. In this regard we mention the work of [4] et al in which they considered the problem:

$$-div(a(x)\,|\nabla u|^{p-2}\nabla u) = f(x,u,\nabla u), \qquad u = 0 \text{ on } \partial\Omega.$$

on a bounded domain $\Omega \subset \mathbb{R}^n$ with $n \geq 2$ and $p \in (1,n)$.In this paper, we investigate the well-posedness of the nonlinear elliptic boundary value problem

$$\begin{cases} Au = -\Delta_p u + div\,(Vu) = R & \text{in } \Omega, \quad p \geq 2, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Δ_p denotes the p-Laplacian operator, $V : \Omega \longrightarrow \mathbb{R}^n$ is a bounded vector field, $V \in L^\infty(\Omega)^n$, and $R{:}\Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is a Carathéodory function satisfying suitable growth and continuity conditions. Such equations naturally arise in models of nonlinear diffusion and convection processes in heterogeneous media.

The existence of weak solutions is established via a priori estimates and fixed-point techniques applied to the compact operator $T = A^{-1}N_R : W_0^{1,p}(\Omega) \longrightarrow W_0^{1,p}(\Omega)$, where $Au = -\Delta_p u + div\,(Vu)$, and N_R denotes the Nemytskii operator associated with the Carathéodory function R, defined by $N_R(x,u) = R(x,u)$.[12] et al. The compactness and continuity of N_R play a crucial role in applying fixed-point theorems to ensure the existence of at least one weak solution. Uniqueness is derived using monotone operator theory by introducing a Carathéodory mapping $b : \Omega \times \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ such that

$$-div\,b(x,u,\nabla u) = -\Delta_p u + div\,(Vu)\,.$$

Problems of this type appear in the modeling of non-Newtonian fluids, nonlinear elasticity, porous media flow, and other nonlinear diffusion–convection phenomena. The present study contributes to the theoretical understanding of such equations and provides new results on the existence and uniqueness of weak solutions.This formulation allows the use of monotone operator theory and variational techniques to prove that the associated operator is strictly monotone under appropriate hypotheses on R and V, there by guaranteeing both existence and uniqueness of the solution to problem (1.1). From an applied perspective, nonlinear elliptic equations involving the p-Laplacian arise in numerous contexts such as non-Newtonian fluid dynamics, nonlinear elasticity, porous medium flow, electro-rheological and thermo-rheological materials, and population dynamics. The convection term $divVu$ introduces significant analytical challenges, as it models transport or drift phenomena in heterogeneous media. Consequently, the study of such equations not only deepens the theoretical understanding of nonlinear PDEs with nonstandard growth but also contributes to the rigorous modeling of complex processes observed in real-world systems.
The paper is organized as follows. In the second section, we define the problem that serves as the focus of our work and we recall some mathematical tools. The Third section will be consecrated to the study of the problem that consists in establishing existence and unicity results.

The organization of this paper is as follows. In Section 2, the considered problem is stated. Sections 3 deals with the solvability of the stated problem. While section 4 gives the uniqueness of the solution of the considered problem.
For the study of this problem, we first recall the following results.

In the sequel, Ω be an open bounded domain of $\mathbb{R}^n$, $n \geqslant 2$ with Lipschitz continuous boundary and $p \geq 1$. We shall use the standard notations:

$$W^{1,p}(\Omega) = \left\{ u \in L^p(\Omega),\ \frac{\partial u}{\partial x_i} \in L^p(\Omega),\ i = 1, ..., n \right\}.$$

equipped with the norm:

$$\|u\|^p_{W^{1,p}(\Omega)} = \|u\|^p_{L^p(\Omega)} + \sum_{i=1}^{i=n} \left\| \frac{\partial u}{\partial x_i} \right\|^p_{L^p(\Omega)},$$

where $\|.\|_{L^p(\Omega)}$ is the usual norm on $L^p(\Omega)$. It is well known that $(W^{1,p}(\Omega);\ \|.\|_{W^{1p,}(\Omega)})$ is separable, reflexive and uniformly convex [13, 14].We need the space $W^{1,p}_0(\Omega)$ the closure of the space $C^\infty_0(\Omega)$ in the space $W^{1,p}(\Omega)$

$$W^{1,p}_0(\Omega) = \left\{ u \in W^{1,p}(\Omega),\ u\,|_{\partial\Omega} = 0 \right\}.$$

The dual space $\left(W^{1,p}_0(\Omega)\right)^*$ will be denoted by $W^{-1,q}(\Omega)$, where $\dfrac{1}{p} + \dfrac{1}{q} = 1$. For each $u \in W^{1,p}(\Omega)$, we put

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, ..., \frac{\partial u}{\partial x_n} \right), \qquad |\nabla u|^2 = \sum_{i=1}^{i=n} \left| \frac{\partial u}{\partial x_i} \right|^2.$$

Definition 1.1. [15] Let X be real Banach space; and let $L : X \longrightarrow X^*$ be an operator Then :

(i) L is called monotone if $\langle Lu - Lv, u - v\rangle \geqslant 0$ for all $u, v \in X$.

(ii) L is called stictly monotone if: $\langle Lu - Lv, u - v\rangle > 0$ for all $u, v \in X$ with $u \neq v$.

(iii) L is called coercive if $\lim\limits_{\|u\| \longrightarrow \infty} \dfrac{\langle Lu, u\rangle}{\|u\|} = +\infty$.

(iv) L is hemicontinuous, i.e., the map $t \longmapsto \langle L(u + tv), w\rangle$ is continuous on $[0, 1]$ for all $u, v, w \in X$.

Definition 1.2. The function $R : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is a Carathéodory function if for each $s \in \mathbb{R}$, the function $x \longmapsto R(x, s)$ is measurable in Ω and for a.e. $x \in \Omega$, the function $s \longmapsto R(x, s)$ is continuous in $\mathbb{R}$.

Theorem 1.3. *[16](Leray−Schauder Principle). Suppose that*

(i) *The operator* $T : X \longrightarrow X$ *is compact, with* X *a Banach space;*

(ii) *There existe an* $r > 0$ *such that if* $x = tT(x)$ *with* $0 < t < 1$ *then* $\|x\| \leqslant r$. *Then the equation* $x = T(x)$ *is solvable.*

2. Problem Statement

In the domain Ω we consider the problem find a solution u such that:

$$\begin{cases} Lu = -\Delta_p u + div(Vu) = R, & \text{in } \Omega, \quad p \geq 2, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where $-\Delta_p$ is the $p-$Laplacian acting from $W_0^{1,p}(\Omega)$ into $W^{-1,p'}(\Omega)$, $\frac{1}{p} + \frac{1}{p'} = 1$ defined by

$$\langle -\Delta_p u, \varphi \rangle = \int_\Omega |\nabla u|^{p-2} \nabla u . \nabla \varphi dx, \quad \forall u, \varphi \in W_0^{1,p}(\Omega), \tag{2.2}$$

The function $V : \Omega \longrightarrow \mathbb{R}^n$ is a vector field $\in L^\infty(\Omega)$ such that

$$divV \geq 0, \tag{2.3}$$

and $R \in W^{-1,q'}(\Omega)$ is a Carathéodory function and it satisfies the growth condition :

$$|R(x,s)| \leqslant a(x)|s|^{q-2} + b(x) \quad \text{p.p } x \in \Omega \text{ and } s \in \mathbb{R}, \tag{2.4}$$

where $a \in L^\infty(\Omega)$, $b \in L^{q'}(\Omega)$, $\frac{1}{q} + \frac{1}{q'} = 1$ and $q \in]1, p^*[$ such that :

$$p^* = \begin{cases} \dfrac{np}{n-p} & \text{if } p < n, \\ +\infty & \text{if } p \geqslant n. \end{cases} \tag{2.5}$$

Proposition 2.1. *If the function* R *is a Carathéodory function, then for each* $u \in M$, *where* M *be the set of all measurable function* $u : \Omega \longrightarrow \mathbb{R}$. *The function* $N_R u : \Omega \longrightarrow \mathbb{R}$ *defined by*

$$(N_R u)(x) = R(x, u(x)) \quad \text{for } x \in \Omega \tag{2.6}$$

is measurable in Ω. *In view of this proposition, a Carathéodory function* $R : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ *defines an operator* $N_R : M \longrightarrow M$, *which is called Nemytskii operator.*

Definition 2.2. An element $u \in W_0^{1,p}(\Omega)$ is a solution of problem (2.1) if

$$-\Delta_p u + div(Vu) = (N_R u), \tag{2.7}$$

in the sens of $W^{-1,p'}(\Omega)$, i.e

$$\int_\Omega |\nabla u|^{p-2} \nabla u . \nabla \varphi dx - \int_\Omega Vu . \nabla \varphi dx = \int_\Omega R(x,u)\varphi dx, \quad \forall u, \varphi \in W_0^{1,p}(\Omega). \tag{2.8}$$

3. Solvability of Problem (2.1)

In this section, the solvability of the problem (2.1) will be proved using the Leray-Schauder principle.

Theorem 3.1. *If the operator* L *is monotone, coercive and hemicontinuous then* L *is surjective.*

Proof: (1) The operator $L : W_0^{1,p}(\Omega) \longrightarrow W^{-1,p'}(\Omega)$ defined by :

$$Lu = -\Delta_p u + div(Vu), \tag{3.1}$$

is bounded.
For $\varphi \in W_0^{1,p}(\Omega)$ we have

$$\langle Lu, \varphi \rangle = \int_\Omega |\nabla u|^{p-2} \nabla u \nabla \varphi dx - \int_\Omega uV \nabla \varphi dx.$$

Using Hölder's inequality we get

$$|\langle Lu, \varphi \rangle| \leq \int_\Omega |\nabla u|^{p-1} . |\nabla \varphi| \, dx + \int_\Omega |u| \, |V| \, |\nabla \varphi| \, dx$$

$$\leq \left(\int_\Omega |\nabla u|^p \, dx \right)^{\frac{p-1}{p}} . \left(\int_\Omega |\nabla \varphi|^p \, dx \right)^{\frac{1}{p}} + \|V\|_{L^\infty(\Omega)} \left(\int_\Omega |u|^p \, dx \right)^{\frac{1}{p}} \left(\int_\Omega |\nabla \varphi|^{p'} \, dx \right)^{\frac{1}{p'}},$$

where $\frac{1}{p} + \frac{1}{p'} = 1$, using the fact that $p \geq 2$ and Pioncaré's inequality, we obtain

$$|\langle Lu, \varphi \rangle| \leq c(\|u\|_{W_0^{1,p}(\Omega)}, \|V\|_{L^\infty(\Omega)}) \|\varphi\|_{W_0^{1,p}(\Omega)}, \tag{3.2}$$

Thus, the operator L is bounded.

(2) If $p \in [2, +\infty[$, then operator L is monotone.
From definition 1.1, the operator L is monotone if and only if $\langle Lu - L\varphi, u - \varphi \rangle \geqslant 0, \ \ \forall u, \varphi \in W_0^{1,p}(\Omega)$, integrating by parts we obtain

$$\langle Lu - L\varphi, u - \varphi \rangle = \langle -\Delta_p u + div(Vu) + \Delta_p \varphi - div(V\varphi), u - \varphi \rangle =$$

$$\int_\Omega \left(|\nabla u|^{p-2} \nabla u - |\nabla \varphi|^{p-2} \nabla \varphi \right) \nabla (u - \varphi) \, dx + \frac{1}{2} \int_\Omega (divV) . (u - \varphi)^2 \, dx, \tag{3.3}$$

Using the fact that

$$\left(|a|^{p-2} a - |b|^{p-2} b \right) (a - b) \geq 2^{-1} \left(|a|^{p-2} + |b|^{p-2} \right) |a - b|^2 \geqslant 2^{2-p} |a - b|^p, \quad \forall a, b \in \mathbb{R}^n, \tag{3.4}$$

then (3.3) becomes

$$\langle Lu - L\varphi, u - \varphi \rangle \geq 2^{2-p} \int_\Omega |\nabla u - \nabla \varphi|^p \, dx + \frac{1}{2} \int_\Omega (divV) . (u - \varphi)^2 \, dx \tag{3.5}$$

According to (2.3) we arrive at

$$\langle Lu - L\varphi, u - \varphi \rangle \geq 0,$$

So, the operator L is monotone.

(3) The operator L is coercive. From (3.5), we have

$$\langle Lu, u \rangle = \langle -\Delta_p u + div(Vu), u \rangle \geq 2^{2-p} \int_\Omega |\nabla u|^p \, dx + \int_\Omega u.V.\nabla u dx. \tag{3.6}$$

According to Hölder's inequality, the second term in the right hand sid of (3.6) can be controlled by

$$\int_\Omega uV.\nabla u dx \leq |\Omega|^{\frac{p-2}{p}} \|V\|_\infty \left(\int_\Omega |u|^p \, dx \right)^{\frac{1}{p}} \left(\int_\Omega |\nabla u|^p \, dx \right)^{\frac{1}{p}},$$

then from (3.6), we get

$$\langle Lu,u\rangle \geq 2^{2-p}\int\limits_{\Omega}|\nabla u|^{p}\,dx - c(\Omega,\|V\|_{\infty})\left(\int\limits_{\Omega}|\nabla u|^{p}\,dx\right)^{\frac{2}{p}}.$$

consequently we have

$$\lim_{\|u\|_{1,p}\longrightarrow+\infty}\frac{\langle Lu,u\rangle}{\|u\|_{1,p}}\longrightarrow+\infty.$$

According to definition (1.1), we conclude that L is coercive.

(4) The operator L is hemicontinuous.
The operator L is hemicontinuous of $W_0^{1,p}(\Omega)$ in $W^{-1,p'}(\Omega)$ if the mapping $t\longmapsto\langle L(u+t_n\varphi),w\rangle$ is continuous in $\mathbb{R}$. Either $u,\varphi,w\in W_0^{1,p}(\Omega)$, $(t_n)_{n\in\mathbb{N}}$ a real sequence that converges to l, we will show that $\langle L(u+t_n\varphi),w\rangle\longrightarrow\langle L(u+l\varphi),w\rangle$.
L is hemicontinuous :

$$\langle L(u+t_n\varphi),w\rangle=\int\limits_{\Omega}|\nabla(u+t_n\varphi)|^{p-2}\nabla(u+t_n\varphi).\nabla w\;dx-\int\limits_{\Omega}(u+t_n\varphi)V.\nabla w\;dx,$$

we have :

$$|\langle L(u+t_n\varphi),w\rangle|\leq\int\limits_{\Omega}|\nabla(u+t_n\varphi)|^{p-1}|\nabla w|\;dx+\int\limits_{\Omega}|u||V||\nabla w|+|t_n|\int\limits_{\Omega}|\varphi||V||\nabla w|\,dx.$$

Using the fact that

$$(a+b)^{p}\leq 2^{p-1}(a^{p}+b^{p}),\ \text{for all } a\geq 0,\ b\geq 0 \text{ and } p\geq 1.$$

we have

$$|\langle L(u+t_n\varphi),w\rangle|\leq 2^{p-1}\int\limits_{\Omega}\left(|\nabla u|^{p-1}+|t_n|^{p-1}|\nabla\varphi|^{p-1}\right)|\nabla w|\;dx+\int\limits_{\Omega}|u||V||\nabla w|+|t_n|\int\limits_{\Omega}|\varphi||V||\nabla w|\,dx. \tag{3.7}$$

Since $t_n\underset{n\longrightarrow\infty}{\longrightarrow}l$, then there exist $M>0$ such that $|t_n|\leq M$, using Hölder's inequality, the right hand side of (3.7) becomes

$$|\langle L(u+t_n\varphi),w\rangle|\leq 2^{p-1}2\frac{1}{p-1}\int\limits_{\Omega}(|\nabla u|^{p}+|M|^{p}|\nabla\varphi|^{p})\left(\int\limits_{\Omega}|\nabla w|^{p}\;dx\right)^{\frac{1}{p}}$$

$$+\|V\|_{\infty}\left(\int\limits_{\Omega}|u|^{p}\,dx\right)^{\frac{1}{p}}\left(\int\limits_{\Omega}|\nabla w|^{p'}\,dx\right)^{\frac{1}{p'}}+M\|V\|_{\infty}\left(\int\limits_{\Omega}|\varphi|^{p}\,dx\right)^{\frac{1}{p}}\left(\int\limits_{\Omega}|\nabla w|^{p'}\,dx\right)^{\frac{1}{p'}}.$$

According to Lebesgue's dominated convergence theorem, the operator L is hemicontinuous. Then the problem $Lu=g(x)$ has a unique solution for $g\in W^{-1,p'}(\Omega)$ i.e the operator L defines one−to−one correspondence between $W_0^{1,p}(\Omega)$ and $W^{-1,p'}(\Omega)$, with inverse L^{-1} monotone, bounded and continuous. Since $q\in(1,q^{*})$, then the imbedding $W_0^{1,p}(\Omega)\hookrightarrow L^{q}$ is compact. Consequently we have the diagram $W_0^{1,p}(\Omega)\overset{I_d}{\hookrightarrow}L^{q}(\Omega)\overset{N_R}{\hookrightarrow}L^{q'}(\Omega)\overset{I_d^{*}}{\hookrightarrow}W_0^{-1,p'}(\Omega)$.

□

Theorem 3.2. *If the Carathéodory function $R:\Omega\times\mathbb{R}\longrightarrow\mathbb{R}$ satisfies the condition (2.4) with $q<p$, then the operator $L^{-1}N_R$ has a fixed point in $W_0^{1,p}(\Omega)$ via Leray-Schauder principe.*

Proof: Using the theorem of Leray-Schauder 1.3, we will show tha the set

$$Z = \left\{ u \in W_0^{1,p}(\Omega) \ / \ \exists \lambda \in [0,1], \text{ such that } u = \lambda L^{-1} N_R u = \lambda T u \right\},$$

is bounded in $W_0^{1,p}(\Omega)$. From (2.7), it is obvious to show that for arbitrary $u \in W_0^{1,p}(\Omega)$, it is obvious that :

$$\|Tu\|^p_{W_0^{1,p}(\Omega)} = \langle -\Delta_p u + div(Vu), Tu \rangle = \langle N_R u, Tu \rangle = \int_\Omega R(x,u)\, Tu \ dx, \tag{3.8}$$

Using (2.4), the right hand side of the previous equality be controlled by

$$\int_\Omega R(x,u)\, Tu \ dx \leq \int_\Omega \left| a(x) |u|^{q-2} + b(x) \right| |Tu| \, dx.$$

Furthermore, for $u \in Z$ i.e. $u = \lambda T u$ with some $\lambda \in [0,1]$, by using Hölder's inequality we get

$$\int_\Omega R(x,u)\, Tu \ dx \leq \int_\Omega |a(x)| \, |\lambda Tu|^{q-2} |Tu| \, dx + \int_\Omega |b(x)| \, |Tu| \, dx$$

$$\leq \lambda^{q-1} \|a(x)\|_{L^\infty(\Omega)} \|Tu\|^{q-1}_{L^{q-1}(\Omega)} + \left(\int_\Omega |b(x)|^{q'} dx \right)^{\frac{1}{q'}} \left(\int_\Omega |Tu|^p dx \right)^{\frac{1}{p}}. \tag{3.9}$$

Combining (3.8) with (3.9) we have :

$$\|Tu\|^p_{1,p} \leqslant K_1 \|Tu\|^{q-1}_{L^q(\Omega)} + K_2 \|Tu\|_{L^q(\Omega)}, \tag{3.10}$$

where

$$K_1 = \lambda^{q-2} \|a(x)\|_{L^\infty(\Omega)},$$

and

$$K_2 = \|b(x)\|_{L^{q'}(\Omega)}.$$

Thus, from (3.10) we obtain

$$\|Tu\|^p_{1,p} - K_1 \|Tu\|^{q-1}_{1,p} - K_2 \|Tu\|_{1,p} \leq 0.$$

Then, there is a constant $d \geqslant 0$ such that

$$\|Tu\|_{1,p} \leqslant d.$$

Since

$$\|u\|_{1,p} = \|\lambda Tu\|_{1,p} = \lambda \|Tu\|_{1,p} \leqslant \lambda d \leqslant d.$$

Then, the bounddeness of Z would de proved, so the problem $-\Delta_p u + div(Vu) = R(x,u)$ has a fixed point $u \in W_0^{1,p}(\Omega)$. □

4. A uniqueness result

In this section we address the problem of the uniqueness of the solution of the problem (2.1):

Theorem 4.1. *For $R \in W^{-1,p'}(\Omega)$, the problem (2.1) has a unique solution $u \in W_0^{1,p}(\Omega)$.*

Proof: Let $b(x,u,\nabla u) = |\nabla u|^{p-2} \nabla u - Vu$. It is obvious to show that

(H1) The function $b : \Omega \times \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a Carathéodory function, i.e. :

a) For each (s,ξ), the function $x \longmapsto b(x,s,\xi)$ is measurable.

b) For a.e. x, the function $(s,\xi) \longmapsto b(x,s,\xi)$ is continuous.

(H2) (Monotonie)

$$\langle b(x,s,\xi) - b(x,s,\eta), \xi - \eta\rangle = \left(|\nabla\xi|^{p-2}\nabla\xi - |\nabla\eta|^{p-2}\nabla\eta\right)(\nabla\xi - \nabla\eta) \geqslant C_p|\xi-\eta|^p.$$

(H3) For almost all $x \in \Omega$ and for each $(x,s,\xi) \in \Omega \times \mathbb{R} \times \mathbb{R}^n$ with $\xi \neq \eta$

$$\langle b(x,s,\xi) - b(x,s,\eta), \xi - \eta\rangle > 0.$$

(H4) According (3.4), there exists $\beta > 0$ such that

$$\langle b(x,s,\xi) - b(x,s,\eta), \xi - \eta\rangle \geqslant \beta|\xi-\eta|^2\left(|\xi|^{p-2} + |\eta|^{p-2}\right).$$

(H5) There exists $C > 0$ and $f \in L^{p'}(\Omega)$ such that :

$$|b(x,s,\xi) - b(x,t,\xi)| \leqslant C|s-t|\left(f(x) + |\xi|^{p-1} + |s|^{p-1} + |t|^{p-1}\right).$$

Let u_1, u_2 be two solutions of problem (2.1) then we have

$$\int_\Omega (b(x,,u_1,\nabla u_1) - b(x,u_1,\nabla u_2))\nabla\varphi dx = \int_\Omega (b(x,,u_2,\nabla u_2) - b(x,u_1,\nabla u_2))\nabla\varphi dx,$$

$T_\varepsilon : \mathbb{R} \longrightarrow \mathbb{R}$ is the truncation function $T_\varepsilon(r) = \max(\min(\varepsilon, r), -\varepsilon)$, for $\varphi = T_\varepsilon u$ and $v = u_1 - u_2$. Since

$$\nabla T_\varepsilon(v) = \nabla\varphi = 1_{A_\varepsilon}\nabla v,$$

where

$$A_\varepsilon = \{x \in \Omega,\ 0 < |u_1 - u_2| \leqslant \varepsilon\},$$

Then, applying **(H4)** and **(H5)** we get:

$$\begin{aligned}
&\beta\int_\Omega |\nabla(u_1-u_2)|^2\left(|\nabla u_1|^{p-2} + |\nabla u_2|^{p-2}\right)dx \\
&\leq C\int_\Omega |u_1-u_2|\left(f(x) + |\nabla u_2|^{p-1} + |u_1|^{p-1} + |u_2|^{p-1}\right)|\nabla(u_1-u_2)|\,dx \\
&\leq C\varepsilon\int_{A_\varepsilon} F|\nabla T_\varepsilon(u_1-u_2)|\,dx,
\end{aligned}$$

where

$$F = \left(f(x) + |\nabla u_2|^{p-1} + |u_1|^{p-1} + |u_2|^{p-1}\right) \in L^{p'}(\Omega).$$

Since

$$\int_{A_\varepsilon} |\nabla(u_1-u_2)|^p dx = \int_\Omega |1_{A_\varepsilon}\nabla(u_1-u_2)|^p dx,$$

and

$$\int_\Omega |\nabla T_\varepsilon(u_1-u_2)|^p dx = \|T_\varepsilon(u_1-u_2)\|^p_{W_0^{1,p}(\Omega)}.$$

Using the fact that

$$\left||z|^{p-2}z - |y|^{p-2}y\right| \leq |z-y|\,|z-y|^{p-2},$$

and (3.4), we obtain

$$\beta \int\limits_{A_\varepsilon} |\nabla (u_1 - u_2)|^p \, dx \leqslant \beta \int\limits_{A_\varepsilon} |\nabla (u_1 - u_2)|^2 \left(|\nabla u_1|^{p-2} + |\nabla u_2|^{p-2} \right) dx,$$

Consequently from the previous inequality we have

$$\beta \left\| T_\varepsilon (u_1 - u_2) \right\|^p_{W_0^{1,p}(\Omega)} \leqslant C\varepsilon \left\| F \right\|_{L^{p'}(\Omega)} \left\| T_\varepsilon (u_1 - u_2) \right\|_{W_0^{1,p}(\Omega)}.$$

The Poincaré (for $p = 1$) and Cauchy-Schwarz inequalities allow us to obtain

$$\begin{aligned} \frac{1}{\varepsilon} \int\limits_{A_\varepsilon} |T_\varepsilon (u_1 - u_2)| \, dx &\leq \frac{C}{\varepsilon} \int\limits_{A_\varepsilon} |\nabla T_\varepsilon (u_1 - u_2)| \, dx \\ &\leq \frac{C}{\varepsilon} \left\| T_\varepsilon (u_1 - u_2) \right\|_{W_0^{1,p}(\Omega)} |A_\varepsilon|^{\frac{1}{p'}} \leq C \, |A_\varepsilon|^{\frac{1}{p'}}. \end{aligned}$$

Gold $\cap_{\varepsilon>0} A_\varepsilon = \emptyset$ and $(A_\varepsilon)_{\varepsilon>0}$ is decreasing when $\varepsilon \longrightarrow 0$. Therefore $|A_\varepsilon| \longrightarrow 0$ when $\varepsilon \longrightarrow 0$.We choose $r > 0$ anything and $\varepsilon < r$, then

$$\begin{aligned} |\{x : \ |u_1 - u_2| \geq r\}| &\leq |\{x : \ |u_1 - u_2| \geq \varepsilon\}| \\ &\leq |\{x : \ |T_\varepsilon (u_1 - u_2)| \geq \varepsilon\}| \\ &\leq \frac{1}{\varepsilon} \int\limits_{\Omega} |T_\varepsilon (u_1 - u_2)| \, dx \longrightarrow 0, \qquad \text{when } \varepsilon \longrightarrow 0. \end{aligned}$$

So, for everything $r > 0$,$|u_1 - u_2| < r$ almost everywhere. Therefore we conclude that $u_1 = u_2$ almost everywhere. □

Proposition 4.2. *If $R(x, s) \geqslant 0$ a.e. $x \in \Omega$ and for all $s \in \mathbb{R}$. If $u \in W_0^{1,p}(\Omega)$ is solution of problem (2.1) then $u \geqslant 0$.*

Proof: Let $u \in W_0^{1,p}(\Omega)$ be a solution of problem (2.1) and let us denoted $\Omega^- = \{x \in \Omega : \ u(x) < 0\}$, $\Omega^+ = \Omega \setminus \Omega^-$ and $u_- = \max\{-u, \ 0\}$, then, it is known that $u_- \in W_0^{1,p}(\Omega)$ and

$$\nabla u_- = \begin{cases} \nabla u & \text{in } \Omega_-, \\ 0 & \text{in } \Omega_+. \end{cases}$$

Integrating by parts over Ω we have

$$\begin{aligned} \langle -\Delta_p u + div(Vu), u_- \rangle &= \int\limits_{\Omega} |\nabla u|^{p-2} \nabla u . \nabla u_- dx_- \int\limits_{\Omega} Vu . \nabla u_- dx \\ &= -\int\limits_{\Omega} |\nabla u|^p \, dx + \int\limits_{\Omega} Vu . \nabla u dx \\ &= -\int\limits_{\Omega_-} |\nabla u|^p \, dx - \int\limits_{\Omega_-} div(V) \frac{u^2}{2} dx = -\int\limits_{\Omega_-} R(x, u) \, u dx \geqslant 0 \end{aligned}$$

Then, we have

$$\int\limits_{\Omega_-} |\nabla u|^p \, dx + \int\limits_{\Omega_-} div(V) \frac{u^2}{2} dx = \int\limits_{\Omega_-} R(x, u) \, u dx \leqslant 0.$$

So

$$\int\limits_{\Omega_-} |\nabla u|^p \, dx + \int\limits_{\Omega_-} div(V) . \frac{u^2}{2} dx = 0.$$

then $u = 0$ everywhere. Wich allows as to achieve the proof. □

References

[1] Cavalheiro A. C., Existence of solutions for the Dirichlet problem of some degenerate semilinear elliptic equations. Applied Mathematics Letters, (2010) Vol. 23, P. 317–321.

[2] Droniou J., Imbert C., Solutions de viscosite et solutions variationnelles pour EDP non lineaires. Universite de Montpellier,

[3] Boccardo L., Gallouet T., Murat F., Mindlin D., Unicite de la solution de certaines equations elliptiques non lineaires. C. R. Acad. Sci. Paris, (1992) Vol. 315, P. 1159–1164.

[4] Motreanu D., Tornatore E., Quasilinear Dirichlet problems with degenerated p-Laplacian and convection term. Mathematics, (2021) Vol. 9, P. 139.

[5] Faraci F., Motreanu D., Puglisi D., Positive solutions of quasi-linear elliptic equations with dependence on the gradient. Calculus of Variations and Partial Differential Equations, (2015) Vol. 54, P. 525–538.

[6] Faria L. F. O., Miyagaki O. H., Motreanu D., Tanaka M., Existence results for nonlinear elliptic equations with Leray–Lions operator and dependence on the gradient. Nonlinear Analysis, (2014) Vol. 96, P. 154–166.

[7] De Figueiredo D., Girardi M., Matzeu M., Semilinear elliptic equations with dependence on the gradient via mountain-pass techniques. Differential Integral Equations, (2004) Vol. 17, P. 119–126.

[8] Franchi B., Serapioni R., Pointwise estimates for a class of strongly degenerate elliptic operators: A geometrical approach. Ann. Scuola Norm. Sup. Pisa, (1987) Vol. 14, P. 527–568.

[9] Marino G., Moser iteration applied to elliptic equations with critical growth on the boundary. Nonlinear Analysis, (2019),

[10] Marino G., Existence and estimates for elliptic equations involving convolution. Computational and Mathematical Methods, (2020),

[11] Averna D., Motreanu D., Tornatore E., Existence and asymptotic properties for quasilinear elliptic equations with gradient dependence. Applied Mathematics Letters, (2016) Vol. 61, P. 102–107.

[12] Dinca G., Jebelean P., Une methode de point fixe pour le p-Laplacien. C. R. Acad. Sci. Paris, (1997) Vol. 324, P. 165–168.

[13] Adams R. A., Sobolev Spaces. Academic Press, (1975).

[14] Zhikov V. V., Weighted Sobolev spaces. Sb. Math. (1998) Vol. 189, Iss. 8, P. 1139–1170.

[15] Zeidler E., Nonlinear Functional Analysis and its Applications. Vol. II/A. Springer-Verlag, New York, (1990).

[16] Zeidler E., Nonlinear Functional Analysis and its Applications. Vol. I: Fixed Point Theorems. Springer-Verlag, New York, (1984).

S. Benaissa,
Department of Mathematics, Mohamed Bachir El Ibrahimi University, BBA, Algeria,
e-mail: soria.benaissa@univ-msila.dz

F. Messelmi,
Department of Mathematics, University of Djelfa, Algeria
e-mail: foudimath@yahoo.fr

A. Merouani,
Department of Mathematics, Ferhat Abbas University,
Sétif, Algeria
e-mail: badri_merouani@yahoo.fr

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 42–48
DOI: 10.29229/uzmj.2026-2-5

Eta quotients of level 20 and weight 1

Bouchikhi A., Mezroui S.

Abstract. We find all the eta quotients in the spaces $M_1\left(\Gamma_0(20), \left(\frac{d}{*}\right)\right)$ with $d = -4, -20$ of modular forms, and we determine their Fourier coefficients, where $\left(\frac{d}{*}\right)$ is the Legendre-Jacobi-Kronecker symbol in the group of Dirichlet characters modulo 20 with values in the rational field $\mathbb{Q}$.

Keywords: Eta quotients, Modular forms, Fourier coefficients of cusp forms, eta function
MSC (2020): 11F11, 11F20, 11F30

1. Introduction

Let $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{Z}$, and $\mathbb{C}$ denote the sets of positive integers, nonnegative integers, integers, and complex numbers, respectively. Let $N \in \mathbb{N}$, $k \in \mathbb{Z}$, and let χ be a Dirichlet character whose modulus divides N. We define the congruence subgroup $\Gamma_0(N)$ of level N by

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z},\ ad - bc = 1,\ c \equiv 0 \pmod{N} \right\}.$$

Let $M_k(\Gamma_0(N), \chi)$ denote the space of modular forms of weight k and character χ on the congruence subgroup $\Gamma_0(N)$. The subspaces of Eisenstein series and cusp forms are denoted by $E_k(\Gamma_0(N), \chi)$ and $S_k(\Gamma_0(N), \chi)$, respectively. It is well-known that

$$M_k(\Gamma_0(N), \chi) = E_k(\Gamma_0(N), \chi) \oplus S_k(\Gamma_0(N), \chi). \tag{1.1}$$

For instance, (see [1, p. 83]).

The Dedekind eta function $\eta(z)$ is a holomorphic function on the upper half-plane

$$\mathbb{H} = \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\},$$

defined by

$$\eta(z) = e^{\pi i z/12} \prod_{n=1}^{\infty} \left(1 - e^{2\pi i n z}\right).$$

A product of the form

$$f(z) = \prod_{1 \le \delta \mid N} \eta^{r_\delta}(\delta z), \tag{1.2}$$

where $r_\delta \in \mathbb{Z}$, not all zero, is called an eta quotient.

Let χ and ψ be Dirichlet characters. For $n \in \mathbb{N}$, we define the generalized divisor sum $\sigma_{(\chi,\psi)}(n)$ by

$$\sigma_{(\chi,\psi)}(n) = \sum_{1 \le m \mid n} \chi(m)\psi(n/m). \tag{1.3}$$

If $n \notin \mathbb{N}$, we set $\sigma_{(\chi,\psi)}(n) = 0$.

Let χ_0 denote the principal character satisfies $\chi_0(m) = 1$ if and only if $(m, N) = 1$, otherwise, $\chi_0(m) = 0$. In particular, we observe that $\sigma_{(\chi_0,\chi_0)}(n)$ coincides with the classical divisor-counting function

$$\sigma_0(n) = \sum_{1 \le m \mid n} 1.$$

We now define two Dirichlet characters modulo 20 by

$$\chi_1(m) = \left(\frac{-4}{m}\right), \qquad \chi_2(m) = \left(\frac{-20}{m}\right), \quad \chi_3(m) = \left(\frac{-5}{m}\right), \qquad (m \in \mathbb{Z}),$$

where $\left(\frac{d}{\cdot}\right)$ denotes the Legendre-Jacobi-Kronecker symbol considered as a Dirichlet character modulo 20 with values in the field of rational numbers with conductor 4, 20 and 5 respectively.

The cusps of $\Gamma_0(N)$ can be represented by rational numbers $\frac{a}{c}$, where $a \in \mathbb{Z}$, $c \in \mathbb{N}$, $c \mid N$, and $\gcd(a, c) = 1$; (see [2, p. 200]).

For the group $\Gamma_0(20)$, a complete set of representatives of the cusps can be chosen as

$$\frac{1}{20}, \quad \frac{1}{10}, \quad \frac{1}{5}, \quad \frac{1}{4}, \quad \frac{1}{2}, \quad \frac{1}{1}. \tag{1.4}$$

Let $f(z)$ be an eta quotient as defined in (1.2). A formula for the order $v_{a/c}(f)$ of f at the cusp $\frac{a}{c}$ (see [2, p. 200]) is given by

$$v_{a/c}(f) = \frac{N}{24\,\gcd(c^2, N)} \sum_{1 \le \delta \mid N} \frac{\gcd(\delta, c)^2\, r_\delta}{\delta}. \tag{1.5}$$

It follows from the dimension formulae [1, Section 6.3] that the only nontrivial modular form spaces of level 20 with trivial cuspidal subspaces are

$$M_1(\Gamma_0(20), \chi_1) \quad \text{and} \quad M_1(\Gamma_0(20), \chi_2).$$

Moreover, we have that

$$\begin{aligned} \dim\big(M_1(\Gamma_0(20), \chi_1)\big) &= \dim\big(E_1(\Gamma_0(20), \chi_1)\big) = 2, \\ \dim\big(M_1(\Gamma_0(20), \chi_2)\big) &= \dim\big(E_1(\Gamma_0(20), \chi_2)\big) = 2. \end{aligned} \tag{1.6}$$

We follow the same principle used in the papers [3, 4]. In this paper, we find all the eta quotients in $M_1(\Gamma_0(20), \chi_1)$ and $M_1(\Gamma_0(20), \chi_2)$, and we determine their Fourier coefficients.

2. Preliminary results

Throughout the remainder of this paper we use the notation $q = q(z) = e^{2\pi i z}$ with $z \in \mathbb{H}$. We define the Eisenstein series $E_{\chi_1,\chi_0}(q)$, $E_{\chi_2,\chi_0}(q)$ and $E_{\chi_2,\chi_1}(q)$ by

$$E_{\chi_1,\chi_0}(q) = \sum_{n=1}^{\infty} \sigma_{(\chi_1,\chi_0)}(n) q^n, \quad E_{\chi_2,\chi_0}(q) = 1 + \sum_{n=1}^{\infty} \sigma_{(\chi_2,\chi_0)}(n) q^n,$$
$$E_{\chi_3,\chi_1}(q) = \sum_{n=1}^{\infty} \sigma_{(\chi_3,\chi_1)}(n) q^n.$$

In view of (2.1) for $N = 20$, we define an eta quotient $f(z)$ by

$$f(z) = \eta^{r_1}(z)\eta^{r_2}(2z)\eta^{r_4}(4z)\eta^{r_5}(5z)\eta^{r_{10}}(10z)\eta^{r_{20}}(20z). \tag{2.1}$$

Theorem 2.1. *Let $f(z) \in M_1(\Gamma_0(20), \chi_1)$ be an eta quotient given by* (2.1)*, and let $f(z) = \sum_{n=0}^{\infty} a_n q^n$ be its Fourier series expansion. Then*

$$f(z) = b_1 E_{\chi_1,\chi_0}(q) + b_2 E_{\chi_1,\chi_0}(q^5)$$

for unique scalars $b_1, b_2 \in \mathbb{C}$, and the Fourier coefficients a_n are given by

$$a_n = b_1 \sigma_{\chi_1,\chi_0}(n) + b_2 \sigma_{\chi_1,\chi_0}(n/5) \text{ for } n \ge 1.$$

Proof: It follows from (1.6) and [1, Theorem 5.9] that the set of Eisenstein series $\{E_{\chi_1,\chi_0}(q),\ E_{\chi_1,\chi_0}(q^5)\}$ is a basis for $M_1(\Gamma_0(20), \chi_1)$. Thus,

$$f(z) = b_1 E_{\chi_1,\chi_0}(q) + b_2 E_{\chi_1,\chi_0}(q^5)$$

for some unique scalars $b_1, b_2, \in \mathbb{C}$, from which the assertion follows by equating the coefficients of q^n on both sides. □

Similarly to Theorem 2.1, we prove the following theorem.

Theorem 2.2. *Let $f(z) \in M_1(\Gamma_0(20), \chi_2)$ be an eta quotient given by (2.1), and let $f(z) = \sum_{n=0}^{\infty} a_n q^n$ be its Fourier series expansion. Then*

$$f(z) = b_1 E_{\chi_2,\chi_0}(q) + b_2 E_{\chi_3,\chi_1}(q)$$

for unique scalars $b_1, b_2 \in \mathbb{C}$, and the Fourier coefficients a_n are given by

$$\begin{aligned} a_n &= b_1 \sigma_{\chi_2,\chi_0}(n) + b_2 \sigma_{\chi_3,\chi_1}(n) \text{ for } n \geq 1, \\ a_0 &= b_1. \end{aligned}$$

We use the following lemma to determine if certain eta quotients are modular forms. See [5, Theorem 5.7, p. 99], [6, Corollary 2.3, p. 37], [7, p. 174], and [8].

Lemma 2.3. *Let $f(z)$ be an eta quotient given by (2.1), and let $k = \frac{1}{2}\sum_{1\leq\delta|N} r_\delta$ and $s = \prod_{1\leq\delta|N} \delta^{r_\delta}$. Suppose that the following conditions are satisfied:*

(i) $\sum_{1\leq\delta|N} \delta \cdot r_\delta \equiv 0 \,(mod\ 24)$,

(ii) $\sum_{1\leq\delta|N} \frac{N}{\delta} \cdot r_\delta \equiv 0 \,(mod\ 24)$,

(iii) $\sum_{1\leq\delta|N} \frac{\gcd(d,\delta)^2 \cdot r_\delta}{\delta} \geq 0$ *for each positive divisor d of N,*

(iv) *k is an integer.*

Then $f(z) \in M_k(\Gamma_0(N), \chi)$, where the character χ is given by

$$\chi(m) = \left(\frac{(-1)^k s}{m}\right).$$

We utilize the following notation for eta-quotients

$$[\delta_1, r_{\delta_1}; \delta_2, r_{\delta_2}; \ldots, \delta_{d-1}, r_{\delta_{d-1}}; \delta_d, r_{\delta_d}] = \prod_{i=1}^{d} \eta^{r_{\delta_i}}(\delta_i z),$$

We take $N = 20$ and $k = 1$ in Lemma 2.3 to obtain the following theorem.

Theorem 2.4. *Let $f(z)$ be an eta quotient given by (2.1), which satisfies the conditions (i)-(iv) in Lemma 2.3 with*

$$r_1 + r_2 + r_4 + r_5 + r_{10} + r_{20} = 2. \tag{2.2}$$

Then $f(z) \in M_1(\Gamma_0(20), \chi)$, where the character χ is determined by

$$f(z) \in \begin{cases} M_1(\Gamma_0(20), \chi_1), & \text{if } r_4 + r_5 + r_{20} \equiv 0 \pmod 2; \\ M_1(\Gamma_0(20), \chi_2), & \text{if } r_2 + r_5 + r_{20} \equiv 0 \pmod 2. \end{cases} \tag{2.3}$$

Proof: For $N = 20$, we have

$$s = \prod_{1\leq\delta|20} \delta^{r_\delta} = 1^{r_1} 2^{r_2} 4^{r_4} 5^{r_5} 10^{r_{10}} 20^{r_{20}} = 2^{r_2+2r_4+r_{10}+2r_{20}} 5^{r_5+r_{10}+r_{20}}.$$

The conditions (i) and (ii) in Lemma 2.3 becomes

$$r_1 + 2r_2 + 4r_4 + 5r_5 + 10r_{10} + 20r_{20} \equiv 0 \pmod{24},$$

$$20r_1 + 10r_2 + 5r_4 + 4r_5 + 2r_{10} + r_{20} \equiv 0 \pmod{24}.$$

Then,

$$r_2 + r_{10} \equiv 0 \pmod 2.$$

The character associated to $f(z)$ is given by

$$\chi(m) = \left(\frac{(-1)^k s}{m}\right) = \left(\frac{-s}{m}\right).$$

Hence,

$$\begin{aligned}
&\chi = \chi_1 \quad \text{if } r_4 + r_5 + r_{20} \equiv 0 \pmod 2,\\
&\chi = \chi_2 \quad \text{if } r_5 + r_{10} + r_{20} \equiv 0 \pmod 2 \quad \text{and } r_2 + r_{10} \equiv 0 \pmod 2.
\end{aligned}$$

These match the given definitions of χ_1 and χ_2, and the result follows. □

3. Main results

Theorem 3.1. *Let $f(z)$ be an eta quotient given by* (2.1). *Then we have $f(z) \in M_1(\Gamma_0(20), \chi_1)$ if and only if*

$$\begin{aligned}
r_1 + 2r_2 + 4r_4 + 5r_5 + 10r_{10} + 20r_{20} &\equiv 0 \pmod{24},\\
20r_1 + 10r_2 + 5r_4 + 4r_5 + 2r_{10} + r_{20} &\equiv 0 \pmod{24},\\
0 \le v_{1/c}(f) < 2 \text{ for } c &= \frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{1}{4}, \frac{1}{2}, \frac{1}{1},\\
r_1 + r_2 + r_4 + r_5 + r_{10} + r_{20} &= 2,\\
r_4 + r_5 + r_{20} &\equiv 0 \pmod 2.
\end{aligned}$$

Proof: Let $f(z) \in M_1(\Gamma_0(20), \chi_1)$ be an eta quotient given by (2.1). By (1.6) we have $\dim(M_1(\Gamma_0(20), \chi_1)) = 2$. We define the eta quotients $f_1(z), f_2(z)$ by

$$\begin{aligned}
f_1(z) &= [-1, 1; 2, 2; 4, -1; 5, 1; 10, 0; 20, 1],\\
f_2(z) &= [1, 1; 2, 0; 4, 1; 5, -1; 10, 2; 20, -1].
\end{aligned}$$

By Lemma 2.3, we have $f_1(z), f_2(z) \in M_1(\Gamma_0(20), \chi_1)$. One can easily see that the set $\{f_1(z), f_2(z)\}$ is linearly independent, and so it is a basis for $M_1(\Gamma_0(20), \chi_1)$. Appealing to (1.4) and (1.5), we have

$v_{1/c}(\cdot)/c$	1	1/2	1/4	1/5	1/10	1/20
f_1	0	1/2	0	1	1/2	1
f_2	1	1/2	1	0	1/2	0

Thus, for any $b_1, b_2 \in \mathbb{C}$ we have

$$v_{1/c}(b_1 f_1 + b_2 f_2 \in \mathbb{N}_0.$$

As $f(z)$ can be expressed as a linear combination of $f_1(z)$ and $f_2(z)$, we have

$$v_{1/c}(f) \in \mathbb{N}_0,$$

from which the second and first assertions follow, respectively. The third assertion follows from [6, Corollary 2.3] and the fifth assertion follows from (2.3). The converse follows from Theorem 2. □

Theorem 3.2. *Let $f(z)$ be an eta quotient given by* (2.1). *Then we have $f(z) \in M_1(\Gamma_0(20), \chi_2)$ if and only if*

$$\begin{aligned}
r_1 + 2r_2 + 4r_4 + 5r_5 + 10r_{10} + 20r_{20} &\equiv 0 \pmod{24},\\
20r_1 + 10r_2 + 5r_4 + 4r_5 + 2r_{10} + r_{20} &\equiv 0 \pmod{24},\\
0 \le v_{1/c}(f) < 2 \text{ for } c &= 1, 2, 4, 5, 10, 20,\\
r_1 + r_2 + r_4 + r_5 + r_{10} + r_{20} &= 2,\\
r_2 + r_5 + r_{20} &\equiv 0 \pmod 2.
\end{aligned}$$

Proof: Let $f(z) \in M_1(\Gamma_0(20), \chi_2)$ be an eta quotient given by (2.1). By (1.6) we have $\dim(M_1(\Gamma_0(20), \chi_2)) = 4$. We define the eta quotients $f_1(z)$ and $f_2(z)$ by

$$f_1(z) = [1, -1; 2, 1; 4, 1; 5, 1; 10, 1; 20, -1],$$
$$f_2(z) = [1, 1; 2, 1; 4, -1; 5, -1; 10, 1; 20, 1].$$

By Lemma 2.3, we have $f_1(z), f_2(z) \in M_1(\Gamma_0(20), \chi_2)$. One can easily see that the set $\{f_1(z), f_2(z)\}$ is a basis for $M_1(\Gamma_0(20), \chi_2)$. By (1.4) and (1.5), we have

$v_{1/c}(\cdot)/c$	1	1/2	1/4	1/5	1/10	1/20
f_1	0	1/2	1	1	1/2	0
f_2	1	1/2	0	0	1/2	1

Then, $v_{1/c}(b_1 f_1 + b_2 f_2 \in \mathbb{N}_0$ for any $b_1, b_2 \in \mathbb{C}$. As $f(z)$ can be expressed as a linear combination of $f_1(z)$ and $f_2(z)$, we have $v_{1/c}(f) \in \mathbb{N}_0$, from which the second and first assertions follow, respectively. The third assertion follows from [6, Corollary 2.3] and the fifth assertion follows from (2.3). The converse follows from Theorem 2. □

There are 4 eta quotients in $M_1(\Gamma_0(20), \chi_1)$ and 3 eta quotients in $M_1(\Gamma_0(20), \chi_2)$. We found all the eta quotients with Sagemath [9] using Theorems 3.1 and 3.2. We then determined their Fourier coefficients using Theorems 2.1 and 2.2. All these eta quotients and their Fourier coefficients are listed in the appendices (Table 1 and 2).

4. Applications

Theorem 4.1. *Let $f(z) \in M_1(\Gamma_0(20), \chi_1)$ with the Fourier series representation*

$$f(z) = \sum_{n=0}^{\infty} a_n q^n.$$

Then For all $n \geq 0$, we have $a_n = 0$ if $n \equiv 3, 7, 11, 15, \pmod{20}$.

Proof: Suppose $n \equiv 3, 7, 11, 15, \pmod{20}$. Then

$$\chi_1(n) = \left(\frac{-4}{n}\right) = \left(\frac{n}{4}\right) = -1.$$

Also, for all positive divisors d of n, we have

$$\chi_1(n/d) = \left(\frac{-4}{n/d}\right) = \left(\frac{-4}{nd}\right) = \left(\frac{-4}{n}\right)\left(\frac{-4}{d}\right) = -\left(\frac{-4}{d}\right) = -\chi_1(d).$$

By pairing $\chi_1(d)$ and $\chi_1(n/d)$ for all $d \mid n$, we obtain

$$\sum_{d|n} \chi_1(d) = 0.$$

The assertion now follows from (1.4), (1.3), and Theorem 2. □

Theorem 4.2. *Let $f(z) \in M_1(\Gamma_0(20), \chi_2)$ with the Fourier series representation*

$$f(z) = \sum_{n=0}^{\infty} a(n) q^n.$$

Then for all $n \geq 0$, we have $a(n) \equiv \pmod{\sigma_{(\chi_3, \chi_1)}(n))}$ if $n \equiv 11, 13, 17, 19 \pmod{20}$.

Proof: Suppose $n \equiv 11, 13, 17, 19 \pmod{20}$. Then

$$\chi_2(n) = \left(\frac{-20}{n}\right) = \left(\frac{n}{20}\right) = -1.$$

Also, for all positive divisors d of n, we have

$$\chi_2(n/d) = \left(\frac{-20}{n/d}\right) = \left(\frac{-20}{nd}\right) = \left(\frac{-20}{n}\right)\left(\frac{-20}{d}\right) = -\left(\frac{-20}{d}\right) = -\chi_2(d).$$

By pairing $\chi_2(d)$ and $\chi_2(n/d)$ for all $d \mid n$, we obtain

$$\sum_{d|n} \chi_2(d) = 0. \tag{4.1}$$

The assertion now follows from (4.1), (1.3), and Theorem 2.2. □

The following corollary follows immediately from Theorem 4.1 and 4.2.

Corollary 4.3. *If an eta quotient $f(z)$ given by* (2.1) *is a modular form of weight* 1 *with the Fourier series representation $f(z) = \sum_{n=0}^{\infty} a_n q^n$, then for all $n \geq 0$ we have*

$$a_n \equiv \begin{cases} 0 \pmod{\sigma_{\chi_3,\chi_1}(n)}, & \text{if } n \equiv 3, 7, 11, 15 \pmod{20},\ r_4 + r_5 + r_{20} \equiv 0 \pmod 2; \\ 0 \pmod{\sigma_{\chi_3,\chi_1}(n)}, & \text{if } n \equiv 11, 13, 17, 19 \pmod{20},\ r_2 + r_5 + r_{20} \equiv 0 \pmod 2. \end{cases}$$

5. Conclusions

In this paper, we have systematically investigated eta quotients of level 20 and weight 1 in the spaces $M_1(\Gamma_0(20), \chi_1)$ and $M_1(\Gamma_0(20), \chi_2)$, where χ_1 and χ_2 are Dirichlet characters modulo 20 associated with the Legendre-Jacobi-Kronecker symbols $\left(\frac{-4}{\cdot}\right)$ and $\left(\frac{-20}{\cdot}\right)$, respectively. Our main contributions include:

(1) Classification of Eta Quotients:
Using the conditions derived from Lemma 2.3, we identified all possible eta quotients in $M_1(\Gamma_0(20), \chi_1)$ and $M_1(\Gamma_0(20), \chi_2)$. Specifically, we found that there are 4 eta quotients in $M_1(\Gamma_0(20), \chi_1)$ and 3 in $M_1(\Gamma_0(20), \chi_2)$, as listed in the appendices (Tables 1 and 2).

(2) Fourier Coefficients:
By expressing these eta quotients in terms of Eisenstein series (Theorems 2.1 and 2.2), we derived explicit formulas for their Fourier coefficients. These results provide a concrete way to compute the coefficients a_n for each modular form.

(3) Vanishing and Congruence Properties: We established vanishing conditions for the Fourier coefficients a_n in certain arithmetic progressions (Theorems 4.1 and 4.2). Specifically, we showed that:

   (a) For forms in $M_1(\Gamma_0(20), \chi_1)$, $a_n = 0$ when $n \equiv 3, 7, 11, 15 \pmod{20}$.

   (b) For forms in $M_1(\Gamma_0(20), \chi_2)$, $a_n \equiv 0 \pmod{\sigma_{(\chi_3,\chi_1)}(n)}$ when $n \equiv 11, 13, 17, 19 \pmod{20}$.

These findings contribute to the broader understanding of modular forms and eta quotients, particularly in cases where the weight is odd and the level is composite. The results may also have applications in number theory, such as in the study of partition functions, quadratic forms, and L-functions.

Acknowledgements. The authors would like to thank the anonymous referee for the valuable suggestions and valuable comments which improved the quality of this paper.

TABLE 1. Eta quotients in $M_1(\Gamma_0(20), \chi_1)$ and their Fourier coefficients

No.	r_1	r_2	r_4	r_5	r_{10}	r_{20}	(b_1, b_2)
1	-4	10	-4	0	0	0	$(4, 0)$
2	-1	2	-1	1	0	1	$(1, 0)$
3	0	0	0	-4	10	-4	$(0, 2)$
4	1	0	1	-1	2	-1	$(-1, 0)$

TABLE 2. Eta quotients in $M_1(\Gamma_0(20), \chi_2)$ and their Fourier coefficients

No.	r_1	r_2	r_4	r_5	r_{10}	r_{20}	(b_1, b_2)
1	-2	5	-2	-2	5	-2	$(1, 1)$
2	-1	1	1	1	1	-1	$(1, 0)$
3	1	1	-1	-1	1	1	$(0, 1)$

REFERENCES

[1] Stein W. A., Modular Forms, A Computational Approach. Amer. Math. Soc., Providence, (2007).

[2] Okamoto A., On expressions of theta series by η-products. Tokyo J. Math. (2011) Vol. 34, P. 319–326.

[3] Alaca A., Alaca S., Aygin Z. S., Eta quotients of level 12 and weight 1. Turkish J. Math. (2019) Vol. 43, Iss. 1, P. 1–8.

[4] Bouchikhi A., Mezroui S., Eta Quotients of Level 18 and Weight 1: Classification and Applications. Bol. Soc. Paran. Mat. (2026) Vol. 44, Iss. 3, P. 1–7.

[5] Kilford L. J. P., Modular Forms: A Classical and Computational Introduction. Imperial College Press, London, (2015).

[6] Kohler G., Eta Products and Theta Series Identities. Springer, New York, (2011).

[7] Gordon B., Sinor D., Multiplicative properties of η-products. Number Theory, Madras 1987 (Proc. Ramanujan Centenary Conf.) Berlin, (2006), P. 173–200.

[8] Ligozat G., Courbes modulaires de genre 1. Bull. Soc. Math. France, (1975) Vol. 43, P. 5–80.

[9] Stein W. A. Sage Mathematics Software (Version 10.4). The Sage Development Team. 2023.

Bouchikhi Abdelmonaim,
Mathematics and Intelligent Systems (MASI),
National School of Applied Sciences of Tangier (ENSAT),
Abdelmalek Essaadi University, Tangier, Morocco
e-mail: bouchikhi.abdelmonaim@gmail.com

Mezroui Soufiane,
Mathematics and Intelligent Systems (MASI),
National School of Applied Sciences of Tangier (ENSAT),
Abdelmalek Essaadi University, Tangier, Morocco
e-mail: smezroui@uae.ac.ma

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 49–58
DOI: 10.29229/uzmj.2026-2-6

Bounds for a class of two-dimensional integrals involving mixed power-type singularities

Chesneau C.

Abstract. This article investigates two-dimensional integrals involving singularities and multiplicative compositions. The main focus is on deriving lower and upper bounds under various assumptions on the primary function, including monotonicity, convexity, and sub-multiplicativity. The results are presented with complete proofs and are intended to motivate further developments in the study of integral inequalities.

Keywords: Two-dimensional integrals, singular integrals, integral inequalities, mixed power-type singularities, bounds
MSC (2020): 26D15, 26A33, 45A05

1. Introduction

Two-dimensional integrals involving singularities frequently arise in the fields of analysis, probability and partial differential equations. The most famous examples are Hilbert-Hardy-type integrals, which are of the following form:

$$\int_0^{+\infty}\int_0^{+\infty} k(x,y)f(x)g(y)dxdy, \tag{1.1}$$

where $k:[0,+\infty)^2\to[0,+\infty)$ is a kernel function, and $f,g:[0,+\infty)\to[0,+\infty)$ are the primary functions of interest. A key aim of this framework is to derive inequalities that bound the integral in Equation (1.1) from above by a constant times the product of two (possibly weighted) integral norms of f and g. Comprehensive overviews of this topic can be found in [1, 2, 3, 4]. Both classical and modern approaches, including techniques for sharp estimates and extensions to various functional settings, are discussed in [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

In this article, we focus on integrals of the following specific form:

$$\int_0^{\beta}\int_0^{\alpha} \frac{1}{x^p+y^q} f(xy)dxdy, \tag{1.2}$$

where $\alpha,\beta>0$ are fixed, $p,q>0$, and $f:[0,\alpha\beta]\to[0,+\infty)$ is the primary function of interest. The integrand in Equation (1.2) exhibits a mixed singularity structure, governed by the interaction between the terms x^p and y^q, combined with a nonlinear dependence on the product xy through the function f. This article aims to establish effective lower and upper bounds for the integral in Equation (1.2) under various assumptions on f, such as monotonicity, convexity, and sub-multiplicativity. Thus, we analyze how the interplay between the singular kernel $k(x,y)=1/(x^p+y^q)$ and the composite function $f(xy)$ influences the behavior of the integral. These results have applications in potential theory, weighted integral inequalities, and probabilistic models involving joint multiplicative structures. All findings are presented with complete proofs and are intended to stimulate further research in the field of integral inequalities.

The rest of the article is organized as follows: Section 2 presents the lower bounds under the monotonicity assumption on f. Section 3 is devoted to the upper bounds under various assumptions on f. Section 4 provides a conclusion.

2. Lower bounds

A lower bound for the integral in Equation (1.2) under the monotonicity assumption on f is given below.

Proposition 2.1. *Let $p, q, \alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a monotonic function with $f(0) < +\infty$ and $f(\alpha\beta) < +\infty$. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \geq \frac{\alpha\beta}{\alpha^p + \beta^q} \min(f(\alpha\beta), f(0)).$$

Proof: For any $x \in [0, \alpha]$ and $y \in [0, \beta]$, we have $x^p + y^q \leq \alpha^p + \beta^q$ and, since f is monotonic, i.e., non-decreasing or non-increasing,

$$f(xy) \geq \min(f(\alpha\beta), f(0)).$$

These inequalities imply that

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \geq \int_0^\beta \int_0^\alpha \frac{1}{\alpha^p + \beta^q} \min(f(\alpha\beta), f(0)) dxdy \\
&= \frac{1}{\alpha^p + \beta^q} \min(f(\alpha\beta), f(0)) \left(\int_0^\alpha dx \right) \left(\int_0^\beta dy \right) \\
&= \frac{\alpha\beta}{\alpha^p + \beta^q} \min(f(\alpha\beta), f(0)).
\end{aligned}$$

This completes the proof of the proposition. □

It is worth noting that if the double integral diverges to $+\infty$, the inequality holds trivially.

The proposition below establishes a more refined lower bound. Note that it is assumed that $p, q \geq 1$.

Proposition 2.2. *Let $p, q \geq 1$, $\alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a monotonic function with $f(0) < +\infty$ and $f(\alpha\beta) < +\infty$. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \geq \min(f(\alpha\beta), f(0)) \max\left(\frac{\alpha}{\beta^{q-1}} \log\left(1 + \frac{\beta^q}{\alpha^p}\right), \frac{\beta}{\alpha^{p-1}} \log\left(1 + \frac{\alpha^p}{\beta^q}\right) \right).$$

Proof: For any $x \in [0, \alpha]$ and $y \in [0, \beta]$, since $p \geq 1$, we have $x^p + y^q \leq \alpha^{p-1}x + \beta^q$ and, since f is monotonic,

$$f(xy) \geq \min(f(\alpha\beta), f(0)).$$

These inequalities imply that

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \geq \int_0^\beta \int_0^\alpha \frac{1}{\alpha^{p-1}x + \beta^q} \min(f(\alpha\beta), f(0)) dxdy \\
&= \min(f(\alpha\beta), f(0)) \left(\int_0^\beta dy \right) \left[\int_0^\alpha \frac{1}{\alpha^{p-1}x + \beta^q} dx \right] \\
&= \min(f(\alpha\beta), f(0)) \beta \left[\frac{1}{\alpha^{p-1}} \log(\alpha^{p-1}x + \beta^q) \right]_{x=0}^{x=\alpha} \\
&= \min(f(\alpha\beta), f(0)) \frac{\beta}{\alpha^{p-1}} \log\left(1 + \frac{\alpha^p}{\beta^q}\right). \qquad (2.1)
\end{aligned}$$

Proceeding in a similar way, for any $x \in [0, \alpha]$ and $y \in [0, \beta]$, since $q \geq 1$, we have $x^p + y^q \leq \alpha^p + \beta^{q-1}y$ and, since f is monotonic,

$$f(xy) \geq \min(f(\alpha\beta), f(0)).$$

These inequalities imply that

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \geq \int_0^\beta \int_0^\alpha \frac{1}{\alpha^p+\beta^{q-1}y} \min(f(\alpha\beta), f(0))dxdy \\
&= \min(f(\alpha\beta), f(0)) \left(\int_0^\alpha dx\right) \left[\int_0^\beta \frac{1}{\alpha^p+\beta^{q-1}y} dy\right] \\
&= \min(f(\alpha\beta), f(0))\alpha \left[\frac{1}{\beta^{q-1}} \log(\alpha^p+\beta^{q-1}y)\right]_{y=0}^{y=\beta} \\
&= \min(f(\alpha\beta), f(0)) \frac{\alpha}{\beta^{q-1}} \log\left(1+\frac{\beta^q}{\alpha^p}\right). \qquad (2.2)
\end{aligned}$$

It follows from Equations (2.1) and (2.2) that

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \geq \min(f(\alpha\beta), f(0)) \max\left(\frac{\alpha}{\beta^{q-1}} \log\left(1+\frac{\beta^q}{\alpha^p}\right), \frac{\beta}{\alpha^{p-1}} \log\left(1+\frac{\alpha^p}{\beta^q}\right)\right).$$

This ends the proof of the proposition. □

Using the logarithmic inequality $\log(1+x) \geq x/(1+x)$ with $x > -1$, we have

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \\
&\geq \min(f(\alpha\beta), f(0)) \max\left(\frac{\alpha}{\beta^{q-1}} \log\left(1+\frac{\beta^q}{\alpha^p}\right), \frac{\beta}{\alpha^{p-1}} \log\left(1+\frac{\alpha^p}{\beta^q}\right)\right) \\
&\geq \min(f(\alpha\beta), f(0)) \max\left(\frac{\alpha}{\beta^{q-1}} \times \frac{\beta^q}{\alpha^p} \times \frac{1}{1+\beta^q/\alpha^p}, \frac{\beta}{\alpha^{p-1}} \times \frac{\alpha^p}{\beta^q} \times \frac{1}{1+\alpha^p/\beta^q}\right) \\
&= \frac{\alpha\beta}{\alpha^p+\beta^q} \min(f(\alpha\beta), f(0)),
\end{aligned}$$

so that Proposition 2.2 implies Proposition 2.1 for the case $p, q \geq 1$.

3. Upper bounds

3.1. **Under the monotonicity assumption.** An upper bound for the integral in Equation (1.2) under the monotonicity assumption on f is presented below. Note that it is assumed that $p, q \in (0, 2)$.

Proposition 3.1. *Let $p, q \in (0, 2)$, $\alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a monotonic function with $f(0) < +\infty$ and $f(\alpha\beta) < +\infty$. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \leq 2\max(f(\alpha\beta), f(0)) \frac{\alpha^{1-p/2}\beta^{1-q/2}}{(2-p)(2-q)}.$$

Proof: Using the basic inequality $a^2+b^2 \geq 2ab$ with $a, b \geq 0$, for any $x \in [0, \alpha]$ and $y \in [0, \beta]$, we have

$$x^p + y^q \geq 2x^{p/2}y^{q/2},$$

and, since f is monotonic,

$$f(xy) \leq \max(f(\alpha\beta), f(0)).$$

These inequalities and standard power primitives together with $p, q \in (0, 2)$ give

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \leq \int_0^\beta \int_0^\alpha \frac{1}{2x^{p/2}y^{q/2}} \max(f(\alpha\beta), f(0))dxdy \\
&= \frac{1}{2}\max(f(\alpha\beta), f(0)) \left(\int_0^\alpha \frac{1}{x^{p/2}} dx\right) \left(\int_0^\beta \frac{1}{y^{q/2}} dy\right) \\
&= \frac{1}{2}\max(f(\alpha\beta), f(0)) \frac{2}{2-p}\alpha^{1-p/2} \times \frac{2}{2-q}\beta^{1-q/2} \\
&= 2\max(f(\alpha\beta), f(0)) \frac{\alpha^{1-p/2}\beta^{1-q/2}}{(2-p)(2-q)}.
\end{aligned}$$

This concludes the proof of the proposition. □

Proposition 3.1 also holds if f is bounded from above, by replacing $\max(f(\alpha\beta), f(0))$ with this upper bound.

The proposition below gives a more refined upper bound under the non-decreasingness assumption on f.

Proposition 3.2. *Let $p, q, \alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a non-decreasing function. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \leq \min\left(\alpha^q \int_0^{\beta\alpha} \frac{1}{x^q} f(x) dx, \beta^p \int_0^{\beta\alpha} \frac{1}{x^p} f(x) dx \right),$$

provided that the integrals in the upper bound converge.

Proof: For any $x \in [0, \alpha]$ and $y \in [0, \beta]$, we have $x^p + y^q \geq x^p$ and, since f is non-decreasing,

$$f(xy) \leq f(\beta x).$$

These inequalities and the change of variables $u = \beta x$ give

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \leq \int_0^\beta \int_0^\alpha \frac{1}{x^p} f(\beta x) dxdy = \left(\int_0^\beta dy \right) \left[\int_0^\alpha \frac{1}{x^p} f(\beta x) dx \right]$$
$$= \beta \int_0^\alpha \frac{1}{x^p} f(\beta x) dx = \beta^p \int_0^{\beta\alpha} \frac{1}{u^p} f(u) du. \tag{3.1}$$

Proceeding in a similar way, for any $x \in [0, \alpha]$ and $y \in [0, \beta]$, we have $x^p + y^q \geq y^q$ and, since f is non-decreasing,

$$f(xy) \leq f(\alpha y).$$

These inequalities and the change of variables $v = \alpha y$ gives

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \leq \int_0^\beta \int_0^\alpha \frac{1}{y^q} f(\alpha y) dxdy = \left(\int_0^\alpha dx \right) \left[\int_0^\beta \frac{1}{y^q} f(\alpha y) dy \right]$$
$$= \alpha \int_0^\beta \frac{1}{y^q} f(\alpha y) dy = \alpha^q \int_0^{\beta\alpha} \frac{1}{v^q} f(v) dv. \tag{3.2}$$

Combining Equations (3.1) and (3.2) and uniformizing the notation, we get

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \leq \min\left(\alpha^q \int_0^{\beta\alpha} \frac{1}{x^q} f(x) dx, \beta^p \int_0^{\beta\alpha} \frac{1}{x^p} f(x) dx \right).$$

This completes the proof of the proposition. □

3.2. **Under the convexity assumption.** The proposition below presents an upper bound for the integral in Equation (1.2) under the convexity assumption on f, i.e., for any $x \in [0, \alpha]$, $y \in [0, \beta]$ and $\lambda \in [0, 1]$, we have $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y)$.

Proposition 3.3. *Let $p > 1$, $q = p/(p - 1)$, $\alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a convex function with $f(0) = 0$ and $f(\alpha\beta) < +\infty$. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \leq \max\left(\frac{1}{p}, \frac{1}{q} \right) f(\alpha\beta).$$

Proof: Using the convexity of f and $f(0)=0$, for any $x\in[0,\alpha]$ and $y\in[0,\beta]$, we have

$$f(xy)=f\left(\frac{xy}{\alpha\beta}\alpha\beta+\left(1-\frac{xy}{\alpha\beta}\right)\times 0\right)\leq\frac{xy}{\alpha\beta}f(\alpha\beta)+\left(1-\frac{xy}{\alpha\beta}\right)f(0)=\frac{xy}{\alpha\beta}f(\alpha\beta).$$

Moreover, for any $x\in[0,\alpha]$ and $y\in[0,\beta]$, the Young product inequality gives

$$xy\leq\frac{1}{p}x^p+\frac{1}{q}y^q\leq\max\left(\frac{1}{p},\frac{1}{q}\right)(x^p+y^q).$$

These inequalities imply that

$$\begin{aligned}
&\int_0^\beta\int_0^\alpha\frac{1}{x^p+y^q}f(xy)dxdy\leq\int_0^\beta\int_0^\alpha\frac{1}{x^p+y^q}\times\frac{xy}{\alpha\beta}f(\alpha\beta)dxdy\\
&=\frac{1}{\alpha\beta}f(\alpha\beta)\int_0^\beta\int_0^\alpha\frac{xy}{x^p+y^q}dxdy\\
&\leq\frac{1}{\alpha\beta}f(\alpha\beta)\int_0^\beta\int_0^\alpha\max\left(\frac{1}{p},\frac{1}{q}\right)\frac{x^p+y^q}{x^p+y^q}dxdy\\
&=\frac{1}{\alpha\beta}f(\alpha\beta)\max\left(\frac{1}{p},\frac{1}{q}\right)\left(\int_0^\alpha dx\right)\left(\int_0^\beta dy\right)\\
&=\frac{1}{\alpha\beta}f(\alpha\beta)\max\left(\frac{1}{p},\frac{1}{q}\right)\alpha\beta=\max\left(\frac{1}{p},\frac{1}{q}\right)f(\alpha\beta).
\end{aligned}$$

This ends the proof of the proposition. □

Note that we have

$$\max\left(\frac{1}{p},\frac{1}{q}\right)=\max\left(\frac{1}{p},\frac{p-1}{p}\right)=\begin{cases}\dfrac{p-1}{p}, & \text{if } p\geq 2,\\ \dfrac{1}{p}, & \text{if } p\in(0,2).\end{cases}$$

A more technical upper bound under the decreasingness assumption on f and convexity assumption on $\log(f)$ is given below.

Proposition 3.4. *Let $p>1$, $q=p/(p-1)$, $\alpha,\beta>0$ and $f:[0,\max(\alpha\beta,\alpha^p,\beta^q)]\to[0,+\infty)$ be a non-decreasing function such that $\log(f)$ is convex. Then we have*

$$\begin{aligned}
&\int_0^\beta\int_0^\alpha\frac{1}{x^p+y^q}f(xy)dxdy\\
&\leq\frac{2}{pq}\sqrt{\int_0^{\alpha^p}x^{2(1/p-1)}\arctan\left[\frac{\beta^{q/2}}{\sqrt{x}}\right]f^{2/p}(x)dx}\sqrt{\int_0^{\beta^q}x^{2(1/q-1)}\arctan\left[\frac{\alpha^{p/2}}{\sqrt{x}}\right]f^{2/q}(x)dx},
\end{aligned}$$

provided that the integrals in the upper bound converge.

Proof: First of all, for any $x\in[0,\alpha]$ and $y\in[0,\beta]$, the Young product inequality gives

$$xy\leq\frac{1}{p}x^p+\frac{1}{q}y^q.$$

Using this inequality, the facts that f is non-decreasing, that $\log(f)$ is convex with $1/p+1/q=1$ and that the exponential function is non-decreasing, we have

$$\begin{aligned}
&f(xy)\leq f\left(\frac{1}{p}x^p+\frac{1}{q}y^q\right)=e^{\log f(x^p/p+y^q/q)}\\
&\leq e^{(1/p)\log f(x^p)+(1/q)\log f(y^q)}=e^{\log[f^{1/p}(x^p)f^{1/q}(y^q)]}=f^{1/p}(x^p)f^{1/q}(y^q).
\end{aligned}$$

Using this inequality and the changes of variables $u = x^p$ and $v = y^q$, we obtain

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \\
&\leq \int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f^{1/p}(x^p) f^{1/q}(y^q) dxdy \\
&= \frac{1}{pq} \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} \left[u^{1/p-1} f^{1/p}(u)\right] \left[v^{1/q-1} f^{1/q}(v)\right] dudv \\
&= \frac{1}{pq} \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} f_\dagger(u) f_\ddagger(v) dudv, \qquad (3.3)
\end{aligned}$$

where

$$f_\dagger(u) = u^{1/p-1} f^{1/p}(u), \quad f_\ddagger(v) = v^{1/q-1} f^{1/q}(v).$$

Let us now bound this last double integral term. By a suitable decomposition of the integrand and the Cauchy-Schwarz integral inequality, we have

$$\int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} f_\dagger(u) f_\ddagger(v) dudv = \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{\sqrt{u+v}} \left(\frac{u}{v}\right)^{1/4} f_\dagger(u) \times \frac{1}{\sqrt{u+v}} \left(\frac{v}{u}\right)^{1/4} f_\ddagger(v) dudv \leq \sqrt{A}\sqrt{B}, \qquad (3.4)$$

where

$$A = \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} \sqrt{\frac{u}{v}} f_\dagger^2(u) dudv$$

and

$$B = \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} \sqrt{\frac{v}{u}} f_\ddagger^2(v) dudv.$$

Let us now express A and B.

Applying the Fubini-Tonelli integral theorem and using a standard arctangent primitive, we have

$$\begin{aligned}
A &= \int_0^{\alpha^p} \sqrt{u} f_\dagger^2(u) \left[\int_0^{\beta^q} \frac{1}{\sqrt{v}(u+v)} dv\right] du \\
&= \int_0^{\alpha^p} \sqrt{u} f_\dagger^2(u) \left[\frac{2}{\sqrt{u}} \arctan\left[\sqrt{\frac{v}{u}}\right]\right]_{v=0}^{v=\beta^q} du \\
&= 2 \int_0^{\alpha^p} f_\dagger^2(u) \arctan\left[\frac{\beta^{q/2}}{\sqrt{u}}\right] du. \qquad (3.5)
\end{aligned}$$

For B, we proceed in a similar way. Applying the Fubini-Tonelli integral theorem and using a standard arctangent primitive, we have

$$\begin{aligned}
B &= \int_0^{\beta^q} \sqrt{v} f_\ddagger^2(v) \left[\int_0^{\alpha^p} \frac{1}{\sqrt{u}(u+v)} du\right] dv \\
&= \int_0^{\beta^q} \sqrt{v} f_\ddagger^2(v) \left[\frac{2}{\sqrt{v}} \arctan\left[\sqrt{\frac{u}{v}}\right]\right]_{u=0}^{u=\alpha^p} dv \\
&= 2 \int_0^{\beta^q} f_\ddagger^2(v) \arctan\left[\frac{\alpha^{p/2}}{\sqrt{v}}\right] dv. \qquad (3.6)
\end{aligned}$$

Combining Equations (3.3), (3.4), (3.5) and (3.6), and using the definitions of $f_\dagger$ and $f_\ddagger$, and uniformizing the notation, we get

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \\
&\leq \frac{1}{pq}\sqrt{2\int_0^{\alpha^p} f_\dagger^2(u)\arctan\left[\frac{\beta^{q/2}}{\sqrt{u}}\right]du}\sqrt{2\int_0^{\beta^q} f_\ddagger^2(v)\arctan\left[\frac{\alpha^{p/2}}{\sqrt{v}}\right]dv} \\
&= \frac{2}{pq}\sqrt{\int_0^{\alpha^p} x^{2(1/p-1)}\arctan\left[\frac{\beta^{q/2}}{\sqrt{x}}\right]f^{2/p}(x)dx}\sqrt{\int_0^{\beta^q} x^{2(1/q-1)}\arctan\left[\frac{\alpha^{p/2}}{\sqrt{x}}\right]f^{2/q}(x)dx}.
\end{aligned}$$

This concludes the proof of the proposition. □

Note that we have

$$\frac{2}{pq} = \frac{2(p-1)}{p^2}.$$

Using the arctangent inequality $\arctan(x) \leq \pi/2$ with $x \in \mathbb{R}$, we have

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \frac{\pi}{pq}\sqrt{\int_0^{\alpha^p} x^{2(1/p-1)}f^{2/p}(x)dx}\sqrt{\int_0^{\beta^q} x^{2(1/q-1)}f^{2/q}(x)dx}.$$

Using the arctangent inequality $\arctan(x) \leq x$ with $x \geq 0$, we also have

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \leq \frac{2}{pq}\alpha^{p/4}\beta^{q/4}\sqrt{\int_0^{\alpha^p} x^{2/p-5/2}f^{2/p}(x)dx}\sqrt{\int_0^{\beta^q} x^{2/q-5/2}f^{2/q}(x)dx}.$$

3.3. **Under the sub-multiplicativity assumption.** The proposition below presents an upper bound for the integral in Equation (1.2) under the sub-multiplicativity assumption on f, i.e., for any $x \in [0,\alpha]$ and $y \in [0,\beta]$, we have $f(xy) \leq f(x)f(y)$.

Proposition 3.5. *Let $p, q, \alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a sub-multiplicative function. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \leq \frac{1}{2}\left[\int_0^\alpha \frac{1}{x^{p/2}} f(x)dx\right]\left[\int_0^\beta \frac{1}{x^{q/2}} f(x)dx\right],$$

provided that the integrals in the upper bound converge.

Proof: Using the basic inequality $a^2 + b^2 \geq 2ab$ with $a, b \geq 0$, for any $x \in [0,\alpha]$ and $y \in [0,\beta]$, we have

$$x^p + y^q \geq 2x^{p/2}y^{q/2}.$$

Using this inequality and the fact that f is sub-multiplicative, and uniformizing the notation, we obtain

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p+y^q} f(xy)dxdy \leq \int_0^\beta \int_0^\alpha \frac{1}{2x^{p/2}y^{q/2}} f(x)f(y)dxdy \\
&= \frac{1}{2}\left[\int_0^\alpha \frac{1}{x^{p/2}} f(x)dx\right]\left[\int_0^\beta \frac{1}{y^{q/2}} f(y)dy\right] = \frac{1}{2}\left[\int_0^\alpha \frac{1}{x^{p/2}} f(x)dx\right]\left[\int_0^\beta \frac{1}{x^{q/2}} f(x)dx\right].
\end{aligned}$$

This completes the proof of the proposition. □

A more technical upper bound under the sub-multiplicativity assumption on f is developed below.

Proposition 3.6. *Let $p, q > 1$, $\alpha, \beta > 0$ and $f : [0, \alpha\beta] \to [0, +\infty)$ be a sub-multiplicative function. Then we have*

$$\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \leq \frac{2}{\sqrt{p}\sqrt{q}} \sqrt{\int_0^\alpha x^{1-p} \arctan\left[\frac{\beta^{q/2}}{x^{p/2}}\right] f^2(x) dx} \sqrt{\int_0^\beta x^{1-q} \arctan\left[\frac{\alpha^{p/2}}{x^{q/2}}\right] f^2(x) dx},$$

provided that the integrals in the upper bound converge.

Proof: Using the sub-multiplicativity of f and the changes of variables $u = x^p$ and $v = y^q$, we obtain

$$\begin{aligned}
&\int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(xy) dxdy \\
&\leq \int_0^\beta \int_0^\alpha \frac{1}{x^p + y^q} f(x) f(y) dxdy \\
&= \frac{1}{pq} \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} \left[u^{1/p-1} f(u^{1/p})\right] \left[v^{1/q-1} f(v^{1/q})\right] dudv \\
&= \frac{1}{pq} \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} f_*(u) f_{**}(v) dudv, \qquad (3.7)
\end{aligned}$$

where

$$f_*(u) = u^{1/p-1} f(u^{1/p}), \quad f_{**}(v) = v^{1/q-1} f(v^{1/q}).$$

Let us now bound this last double integral term, proceeding similarly to the proof of Proposition 3.4. By a suitable decomposition of the integrand and the Cauchy-Schwarz integral inequality, we have

$$\begin{aligned}
&\int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} f_*(u) f_{**}(v) dudv \\
&= \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{\sqrt{u+v}} \left(\frac{u}{v}\right)^{1/4} f_*(u) \times \frac{1}{\sqrt{u+v}} \left(\frac{v}{u}\right)^{1/4} f_{**}(v) dudv \\
&\leq \sqrt{C}\sqrt{D}, \qquad (3.8)
\end{aligned}$$

where

$$C = \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} \sqrt{\frac{u}{v}} f_*^2(u) dudv$$

and

$$D = \int_0^{\beta^q} \int_0^{\alpha^p} \frac{1}{u+v} \sqrt{\frac{v}{u}} f_{**}^2(v) dudv.$$

Let us now express C and D.

Applying the Fubini-Tonelli integral theorem and using a standard arctangent primitive, we have

$$\begin{aligned}
C &= \int_0^{\alpha^p} \sqrt{u} f_*^2(u) \left[\int_0^{\beta^q} \frac{1}{\sqrt{v}(u+v)} dv\right] du \\
&= \int_0^{\alpha^p} \sqrt{u} f_*^2(u) \left[\frac{2}{\sqrt{u}} \arctan\left[\sqrt{\frac{v}{u}}\right]\right]_{v=0}^{v=\beta^q} du \\
&= 2 \int_0^{\alpha^p} f_*^2(u) \arctan\left[\frac{\beta^{q/2}}{\sqrt{u}}\right] du. \qquad (3.9)
\end{aligned}$$

For D, we proceed in a similar way. Applying the Fubini-Tonelli integral theorem and using a standard arctangent primitive, we have

$$\begin{aligned} D &= \int_0^{\beta^q} \sqrt{v} f_{**}^2(v) \left[\int_0^{\alpha^p} \frac{1}{\sqrt{u}(u+v)} du \right] dv \\ &= \int_0^{\beta^q} \sqrt{v} f_{**}^2(v) \left[\frac{2}{\sqrt{v}} \arctan \left[\sqrt{\frac{u}{v}} \right] \right]_{u=0}^{u=\alpha^p} dv \\ &= 2 \int_0^{\beta^q} f_{**}^2(v) \arctan \left[\frac{\alpha^{p/2}}{\sqrt{v}} \right] dv. \end{aligned} \tag{3.10}$$

Combining Equations (3.7), (3.8), (3.9) and (3.10), and using the definitions of f_* and f_{**} and the changes of variables $u = x^p$ and $v = x^q$, we get

$$\begin{aligned} &\int_0^{\beta} \int_0^{\alpha} \frac{1}{x^p + y^q} f(xy) dx dy \\ &\leq \frac{1}{pq} \sqrt{2 \int_0^{\alpha^p} f_*^2(u) \arctan \left[\frac{\beta^{q/2}}{\sqrt{u}} \right] du} \sqrt{2 \int_0^{\beta^q} f_{**}^2(v) \arctan \left[\frac{\alpha^{p/2}}{\sqrt{v}} \right] dv} \\ &= \frac{2}{pq} \sqrt{\int_0^{\alpha^p} u^{2(1/p-1)} \arctan \left[\frac{\beta^{q/2}}{\sqrt{u}} \right] f^2(u^{1/p}) du} \\ &\times \sqrt{\int_0^{\beta^q} v^{2(1/q-1)} \arctan \left[\frac{\alpha^{p/2}}{\sqrt{v}} \right] f^2(v^{1/q}) dv} \\ &= \frac{2}{\sqrt{p}\sqrt{q}} \sqrt{\int_0^{\alpha} x^{1-p} \arctan \left[\frac{\beta^{q/2}}{x^{p/2}} \right] f^2(x) dx} \\ &\times \sqrt{\int_0^{\beta} x^{1-q} \arctan \left[\frac{\alpha^{p/2}}{x^{q/2}} \right] f^2(x) dx}. \end{aligned}$$

This ends the proof of the proposition. □

Note that we have

$$\frac{2}{\sqrt{p}\sqrt{q}} = \frac{2\sqrt{p-1}}{p}.$$

Using the arctangent inequality $\arctan(x) \leq \pi/2$ with $x \in \mathbb{R}$, we have

$$\begin{aligned} &\int_0^{\beta} \int_0^{\alpha} \frac{1}{x^p + y^q} f(xy) dx dy \\ &\leq \frac{\pi}{\sqrt{p}\sqrt{q}} \sqrt{\int_0^{\alpha} x^{1-p} f^2(x) dx} \sqrt{\int_0^{\beta} x^{1-q} f^2(x) dx}. \end{aligned}$$

Using the arctangent inequality $\arctan(x) \leq x$ with $x \geq 0$, we also have

$$\begin{aligned} &\int_0^{\beta} \int_0^{\alpha} \frac{1}{x^p + y^q} f(xy) dx dy \\ &\leq \frac{2}{\sqrt{p}\sqrt{q}} \alpha^{p/4} \beta^{q/4} \sqrt{\int_0^{\alpha} x^{1-3p/2} f^2(x) dx} \sqrt{\int_0^{\beta} x^{1-3q/2} f^2(x) dx}. \end{aligned}$$

4. Conclusion

This article provides a detailed analysis of a class of two-dimensional integrals involving mixed singularities and multiplicative compositions. By establishing precise lower and upper bounds on

the integrand function under various assumptions, new insights into the behavior of such integrals are revealed. The methods and results presented here could be applied to related areas of analysis, probability, and mathematical physics. Further generalizations, including multidimensional extensions and alternative singular integrands, could be explored in future research.

References

[1] Hardy G. H., Littlewood J. E., Polya G., Inequalities. Cambridge University Press, Cambridge, (1934).

[2] Chen Q., Yang B. C., A survey on the study of Hilbert-type inequalities. J. Inequal. Appl. (2015) Vol. 2015, P. 1–29.

[3] Yang B. C., Hilbert-Type Integral Inequalities. Bentham Science Publishers, The United Arab Emirates, (2009).

[4] Yang B. C., The Norm of Operator and Hilbert-Type Inequalities. Science Press, Beijing, (2009).

[5] Sun B., Best generalization of a Hilbert type inequality. J. Ineq. Pure Appl. Math. (2006) Vol. 7, P. 1–7.

[6] Li Y., He B., On inequalities of Hilbert's type. Bull. Aust. Math. Soc. (2007) Vol. 76, Iss. 1, P. 1–13.

[7] Li Y., Wu J., He B., A new Hilbert-type integral inequality and the equivalent form. Int. J. Math. Math. Sci. (2006) Vol. 8, P. 1–6.

[8] Azar L. E., On some extensions of Hardy-Hilbert's inequality and applications. J. Ineq. Appl. (2008) Vol. 2008, P. 1–14.

[9] Saglam A., Yildirim H., Sarikaya M. Z., Generalization of Hardy-Hilbert's inequality and applications. Kyungpook Math. J. (2010) Vol. 50, P. 131–152.

[10] Sulaiman W. T., New Hardy-Hilbert's-type integral inequalities. Int. Math. Forum, (2008) Vol. 3, P. 2139–2147.

[11] Sulaiman W. T., Hardy-Hilbert's integral inequality in new kinds. Math. Commun. (2010) Vol. 15, P. 453–461.

[12] Sulaiman W. T., New kinds of Hardy-Hilbert's integral inequalities. Appl. Math. Lett. (2010) Vol. 23, P. 361–365.

[13] Sulaiman W. T., An extension of Hardy-Hilbert's integral inequality. Afr. Diaspora J. Math. (2010) Vol. 10, P. 66–71.

[14] Sarikaya M. Z., Bingol M. S., Recent developments of integral inequalities of the Hardy-Hilbert type. Turkish J. Ineq. (2024) Vol. 8, P. 43–54.

[15] Chesneau C., Refining and extending two special Hardy-Hilbert-type integral inequalities. Ann. Math. Comp. Sci. (2025) Vol. 28, P. 21–45.

Chesneau C.,
Department of Mathematics, LMNO, University of Caen-Normandie, Caen, France.
e-mail: christophe.chesneau@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 59–67
DOI: 10.29229/uzmj.2026-2-7

Inverse problem for a system of mixed-type second-order partial differential equations

Fayazov K., Juraeva D.

Abstract. This paper investigates an inverse problem for a system of mixed-type second-order differential equations with a variable time direction. The problem involves finding the unknown right-hand side functions from the solution of the system under given boundary and initial conditions. The method of separation of variables is applied to reduce the problem to a spectral problem, in which eigenvalues and eigenfunctions are determined. The obtained spectral problem is shown to coincide with previously established spectral problems. The orthogonality of the eigenfunctions is used to derive explicit formulas for the coefficients. The existence and uniqueness of the solution are established using the properties of the eigenfunction system and the Hilbert-Schmidt theorem.

Keywords: Inverse problem, mixed-type differential equations, separation of variables, spectral problem, eigenvalues, eigenfunctions, orthogonality, Hilbert-Schmidt theorem
MSC (2020): 35R30

1. Introduction

Boundary value problems for parabolic equations with variable time direction have been extensively investigated (see, for example, [1]-[2]). Recently, the complexity of formulating inverse problems for these equations has grown, driven by the need to model and control thermal physics and continuum mechanics processes.

The first works on parabolic equations with a changing time direction were by M. Gevrey [1]. Boundary-value problems for such equations and systems were explored by H. Levine [3], K. S. Fayazov, Y. K. Khudayberganov [4], K. S. Fayazov, Y. K. Khudayberganov and S. Pyatkov [5], I. O. Khajiev and Y. K. Khudayberganov [6], K. S. Fayazov, I. O. Khajiev [7], among others. Inverse problems for parabolic equations have been studied quite well by A. I. Kozhanov [8], I. A. Kaliev, M. F. Mugafarov, O. V. Fattahova [2], A. Bubnov [9] and others.

This paper investigates a more general system of equations, with the method used applicable to other systems. Although much progress has been made in inverse problems for parabolic equations [8]-[9], this work's unique contribution is its focus on inverse problems with a time direction that varies.

These equations have widespread applications, including heat transfer in inhomogeneous media, filtration flow interactions, mass transfer near aircraft surfaces, and complex viscous fluid flow. They also have relevance in transport theory, electron scattering [10, 11], plasma theory [12], hydrodynamics (filtration and boundary layer theory) [13, 14, 15], kinetic theory, stochastic processes [16], and diffusion processes [17, 18].

Thus, these problems are important in mathematical modeling, making the task highly relevant as both an applied and mathematical physics problem.

2. Statement of the problem

Let $\Omega = \Omega_- \cup \Omega_+$, where $\Omega_- = (-1,0)\times(0,T)$, $\Omega_+ = (0,1)\times(0,T)$. $u(x,t)$, $v(x,t)$ are solutions in the domain Ω of the system of equations

$$\begin{cases} u_t(x,t) - \operatorname{sign}(x) u_{xx}(x,t) + a_1 u(x,t) + b_1 v(x,t) = f(x)\,, \\ v_t(x,t) - \operatorname{sign}(x) v_{xx}(x,t) + a_2 v(x,t) + b_2 u(x,t) = g(x). \end{cases} \tag{2.1}$$

Find functions $v(x,t)$, $u(x,t)$ and $f(x)$, $g(x)$ satisfying the system of equations (2.1), boundary

$$u_x(\pm 1,t) = 0,\ v_x(\pm 1,t) = 0,\ 0 \le t \le T \tag{2.2}$$

initial

$$u(x,0)=\varphi(x)=\begin{cases}\varphi_1(x), & -1\le x\le 0,\\ \varphi_2(x), & 0\le x\le 1,\end{cases}$$
$$v(x,0)=\phi(x)=\begin{cases}\phi_1(x), & -1\le x\le 0,\\ \phi_2(x), & 0\le x\le 1,\end{cases} \tag{2.3}$$

final

$$u(x,T)=\psi(x)=\begin{cases}\psi_1(x), & -1\le x\le 0,\\ \psi_2(x), & 0\le x\le 1,\end{cases}$$
$$v(x,T)=\gamma(x)=\begin{cases}\gamma_1(x), & -1\le x\le 0,\\ \gamma_2(x), & 0\le x\le 1,\end{cases} \tag{2.4}$$

and gluing

$$\begin{aligned}&u(-0,t)=u(+0,t),\ u_x(-0,t)=-u_x(+0,t),\\ &v(-0,t)=v(+0,t),\ v_x(-0,t)=-v_x(+0,t),\end{aligned}\quad 0\le t\le T. \tag{2.5}$$

conditions, where $\varphi(x),\psi(x),\phi(x),\gamma(x)$ are sufficiently smooth given functions, a_i and b_i are any constants, $b_2\neq 0$, $(a_1-a_2)^2+4b_1b_2>0$.

In this formulation, $f(x),g(x)$ are the unknown right-hand side of the system (2.1). The boundary conditions of the second kind are conditions (2.2) for the corresponding parabolic equations in the domains Ω_- and Ω_+. The initial conditions are specified using the functions $\varphi_2(x)$, $\psi_1(x)$, $\phi_2(x)$, and $\gamma_1(x)$ the final redefinition conditions necessary to find the functions $f(x),g(x)$ are specified using the functions $\varphi_1(x)$, $\psi_2(x)$, $\phi_1(x)$, and $\gamma_2(x)$.

3. Auxiliary statements

In order to study problem (2.1) - (2.5) we make the following substitution

$$\begin{aligned}u(x,t)&=\frac{a_1+k_2}{b_2(k_1-k_2)}\omega_1(x,t)+\frac{a_1+k_1}{b_2(k_1-k_2)}\omega_2(x,t),\\ v(x,t)&=\frac{1}{k_1-k_2}\omega_1(x,t)+\frac{1}{k_1-k_2}\omega_2(x,t),\end{aligned} \tag{3.1}$$

where k_1 and k_2 are the roots of the quadratic equation.

$$k^2+(a_1+a_2)k+a_1a_2-b_1b_2=0.$$

Then we come to the following problems regarding the functions $\omega_i(x,t),i=1,2$.

Problem 1. Find a function $\omega_i(x,t)$ satisfying the equation

$$\omega_{it}(x,t)-\operatorname{sign}(x)\omega_{ixx}(x,t)-k_i\omega_i(x,t)=\bar f_i(x), \tag{3.2}$$

in the domain $\Omega=\Omega_-\cup\Omega_+$ and the following conditions:

$$\omega_{ix}(\pm1,t)=0,\ 0\le t\le T, \tag{3.3}$$

$$\omega_i(x,0)=\bar\varphi_i(x)=\begin{cases}\bar\varphi_{i1}(x), & -1\le x\le 0,\\ \bar\varphi_{i2}(x), & 0\le x\le 1,\end{cases} \tag{3.4}$$

$$\omega_i(x,T)=\bar\psi_i(x)=\begin{cases}\bar\psi_{i1}(x), & -1\le x\le 0,\\ \bar\psi_{i2}(x), & 0\le x\le 1,\end{cases} \tag{3.5}$$

$$\omega_i(-0,t)=\omega_i(+0,t),\quad \omega_{ix}(-0,t)=-\omega_{ix}(+0,t), \tag{3.6}$$

where

$$\begin{aligned}&\bar\varphi_1(x)=(a_1+k_1)\phi(x)-b_2\varphi(x), && \bar\varphi_2(x)=b_2\varphi(x)-(a_1+k_2)\phi(x),\\ &\bar\psi_1(x)=(a_1+k_1)\gamma(x)-b_2\psi(x), && \bar\psi_2(x)=b_2\psi(x)-(a_1+k_2)\gamma(x),\\ &\bar f_1(x)=(a_1+k_1)g(x)-b_2f(x), && \bar f_2(x)=b_2f(x)-(a_1+k_2)g(x).\end{aligned}$$

4. Application of the method of separation of variables

The method of studying the inverse problem in this work is based on the exclusion of the function $\bar{f}_i(x)$ from equation (3.2) using differentiation with respect to time and the transition to the problem for the function

$$\xi_i(x,t) = \omega_{it}(x,t)\,.$$

From equation (3.2) and conditions (3.3) - (3.6) we obtain the following initial boundary value problem. Find the functions $\xi_i(x,t)$ and $\bar{f}_i(x)$ satisfying the equation in the domain Ω

$$\xi_{it}(x,t) - \operatorname{sign}(x)\xi_{ixx}(x,t) - k_i\xi_i(x,t) = 0, \tag{4.1}$$

boundary conditions

$$\xi_{ix}(\pm 1,t) = 0, \quad 0 \le t \le T, \tag{4.2}$$

initial and final conditions

$$\xi_i(x,0) = \bar{f}_i(x) + \operatorname{sign}(x)\bar{\varphi}_i''(x) + k_i\bar{\varphi}_i, \quad x \in (-1,0)\cup(0,1), \tag{4.3}$$

$$\xi_i(x,T) = \bar{f}_i(x) + \operatorname{sign}(x)\bar{\psi}_i''(x) + k_i\bar{\psi}_i, \quad x \in (-1,0)\cup(0,1), \tag{4.4}$$

gluing conditions

$$\xi_i(-0,t) = \xi_i(+0,t), \quad \xi_{ix}(-0,t) = -\xi_{ix}(+0,t), \quad 0 \le t \le T. \tag{4.5}$$

We also use the separable method to find a solution to the (4.1) - (4.5) problem. As a result, we arrive at the following spectral problem.

Spectral problem. Find the values of the parameter λ such that the spectral problem

$$\begin{cases} \operatorname{sign}(x)X''(x) - \lambda X(x) = 0, \\ X'(-1) = X'(1) = 0, \\ X(-0) = X(+0), \\ X'(-0) = -X'(+0) \end{cases} \tag{4.6}$$

has a nontrivial solution.

(4.6) the spectral problem has positive λ_n^+, λ_n^- negative and zero eigenvalues, which are the roots of the transcendental equation $\tan\sqrt{\lambda} + \tanh\sqrt{\lambda} = 0$ for $\forall n \in N$ .The eigenfunctions corresponding to each eigenvalue are described in the form

$$X_n^+(x) = \begin{cases} \frac{\cos\sqrt{\lambda_n^+}(1+x)}{\cos\sqrt{\lambda_n^+}}, & x < 0, \\ \frac{\cosh\sqrt{\lambda_n^+}(1-x)}{\cosh\sqrt{\lambda_n^+}}, & x > 0, \end{cases} \quad X_n^-(x) = \begin{cases} \frac{\cosh\sqrt{-\lambda_n^-}(x+1)}{\cosh\sqrt{-\lambda_n^-}}, & x < 0, \\ \frac{\cos\sqrt{-\lambda_n^-}(1-x)}{\cos\sqrt{-\lambda_n^-}}, & x > 0, \end{cases}$$

$$X_0(x) = \begin{cases} \frac{1}{\sqrt{2}}, & x < 0, \\ \frac{1}{\sqrt{2}}, & x > 0. \end{cases} \tag{4.7}$$

From the results of [18], eigenfunctions of the spectral problem (4.7) are normalized in $L_2(-1;0)\cup(0;1)$ space based on the Hilbert-Schmidt theorem and constitute the Riess basis.

We find the function $T_i(t)$ from the equation

$$T'_i(t) = (\lambda + k_i)T_i(t).$$

Using the above information we get

$$\xi_i(x,0) = \bar{f}_i(x) + \operatorname{sign}(x)\bar{\varphi''}_i(x) + k_i\bar{\varphi}_i,$$

$$T_{i0}X_0(x) + \sum_{n=1}^{\infty} T_{in}^-(0)X_n^-(x) + \sum_{n=1}^{\infty} T_{in}^+(0)X_n^+(x) = \bar{f}_i(x) + \operatorname{sign}(x)\bar{\varphi''}_i(x) + k_i\bar{\varphi}_i,$$

$$T_{in}^{\pm}(0) = \left(\bar{f}_i(x) + \text{sign}(x)\bar{\varphi''}_i(x) + k_i\bar{\varphi}_i, X_n^{\pm}(x)\right) = A_{in}^{\pm}.$$

Using the information at $t = T$, we can write

$$\xi_i(x,T) = \bar{f}_i(x) + \text{sign}(x)\bar{\psi''}_i(x) + k_i\bar{\psi}_i\,,$$

$$T_{i0}X_0(x) + \sum_{n=1}^{\infty} T_n^-(T)X_n^-(x) + \sum_{n=1}^{\infty} T_{in}^+(T)X_n^+(x) = \bar{f}_i(x) + \text{sign}(x)\bar{\psi''}_i(x) + k_i\bar{\psi}_i\,,$$

$$T_{in}^{\pm}(T) = \left(\bar{f}_i(x) + \text{sign}(x)\bar{\psi''}_i(x) + k_i\bar{\psi}_i, X_n^{\pm}(x)\right) = B_{in}^{\pm}\,.$$

1) If $\lambda_n^+ > 0$, then

$$T_{in}^+(t) = C_{in}e^{(\lambda_n^+ + k_i)t},$$
$$T_{in}^+(0) = A_{in}^+,$$
$$T_{in}^+(T) = B_{in}^+,$$
$$T_{in}^+(t) = B_{in}^+e^{-(\lambda_n^+ + k_i)(T-t)}.$$

2) If $\lambda_n^- < 0$, then

$$T_{in}^-(t) = C_{in}e^{(\lambda_n^- + k_i)t},$$
$$T_{in}^-(0) = A_{in}^-,$$
$$T_{in}^-(T) = B_{in}^-,$$
$$T_{in}^-(t) = A_{in}^-e^{(\lambda_n^- + k_i)t}.$$

3) If $\lambda_0 = 0$, then

$$T_{i0}(t) = C_ie^{k_it}.$$

The function $\xi_i(x,t)$ can be presented in the form

$$\xi_i(x,t) = C_ie^{k_it}X_0(x) + \sum_{n=1}^{\infty} A_{in}^-e^{(\lambda_n^- + k_i)t}X_n^-(x) + \sum_{n=1}^{\infty} B_{in}^+e^{-(\lambda_n^+ + k_i)(T-t)}X_n^+(x) \tag{4.8}$$

and we need to find the coefficients A_{in}^-, B_{in}^+, C_i.

To find the coefficients A_{in}^-, B_{in}^+, C_i we use conditions (4.3) and (4.4)

$$\bar{f}_i(x) = \xi_i(x,0) - \text{sign}(x)\bar{\varphi''}_i(x) - k_i\bar{\varphi}_i = [C_iX_0(x) +$$
$$\sum_{n=1}^{\infty} A_{in}^-X_n^-(x) + \sum_{n=1}^{\infty} B_{in}^+e^{-(\lambda_n^+ + k_i)T}X_n^+(x)\Bigg] - \text{sign}(x)\bar{\varphi''}_i(x) - k_i\bar{\varphi}_i(x)\,,$$

$$\bar{f}_i(x) = \xi_i(x,T) - \text{sign}(x)\bar{\psi''}_i(x) - k_i\bar{\psi}_i = \Bigg[C_ie^{k_iT}X_0(x) + \sum_{n=1}^{\infty} A_{in}^-e^{(\lambda_n^- + k_i)T}X_n^-(x)$$
$$+ \sum_{n=1}^{\infty} B_{in}^+X_n^+(x)\Bigg] - \text{sign}(x)\bar{\psi''}_i(x) - k_i\bar{\psi}_i(x)\,.$$

From the last two equalities we have

$$C_i\left(1 - e^{k_iT}\right)X_0(x) + \sum_{n=1}^{\infty} A_{in}^-\left(1 - e^{(\lambda_n^- + k_i)T}\right)X_n^-(x) +$$
$$\sum_{n=1}^{\infty} B_{in}^+\left(e^{-(\lambda_n^+ + k_i)T} - 1\right)X_n^+(x) = \text{sign}(x)\left(\bar{\varphi''}_i(x) - \bar{\psi''}_i(x)\right) + k_i\left(\bar{\varphi}_i(x) - \bar{\psi}_i(x)\right). \tag{4.9}$$

We scalar multiply the left and right parts of equality (4.9) by X_n^- to find the coefficient A_{in}, by X_0 to find the coefficient C_i and by X_n^+ to find the coefficient B_{in}. Taking into account the orthogonality of the system of eigenfunctions, we obtain the following formulas:

$$A_{in}^- = \frac{\bar{\varphi}_{in}^{-(2)} - \bar{\psi}_{in}^{-(2)} + k_i(\bar{\varphi}_{in}^- - \bar{\psi}_{in}^-)}{1 - e^{(\lambda_n^- + k_i)T}}, \tag{4.10}$$

$$B_{in}^+ = \frac{\bar{\varphi}_{in}^{+(2)} - \bar{\psi}_{in}^{+(2)} + k_i(\bar{\varphi}_{in}^+ - \bar{\psi}_{in}^+)}{e^{(\lambda_n^+ + k_i)T} - 1}, \tag{4.11}$$

$$C_i = \frac{\bar{\varphi}_{i0}^{(2)} - \bar{\psi}_{i0}^{(2)} + k_i(\bar{\varphi}_{i0} - \bar{\psi}_{i0})}{1 - e^{k_i T}}, \tag{4.12}$$

where $\bar{\varphi}_{in}^{\pm(2)} = (\mathrm{sign}x \cdot \bar{\varphi''}_i(x), X_n^\pm(x))$, $\bar{\psi}_{in}^{\pm(2)} = (\mathrm{sign}x \cdot \bar{\psi''}_i(x), X_n^\pm(x))$, $\bar{\varphi}_{i0}^{(2)} = (\mathrm{sign}x \cdot \bar{\varphi''}_i(x), X_0(x))$, $\bar{\psi}_{i0}^{(2)} = (\mathrm{sign}x \cdot \bar{\psi''}_i(x), X_0(x))$, $\bar{\varphi}_{in}^\pm = (\bar{\varphi}_i(x), X_n^\pm(x))$, $\bar{\psi}_{in}^\pm = (\bar{\psi}_i(x), X_n^\pm(x))$, $\bar{\varphi}_{i0} = (\bar{\varphi}_i(x), X_0(x))$, $\bar{\psi}_{i0} = (\bar{\psi}_i(x), X_0(x))$.

5. The theorem of existence and uniqueness of a solution to the problem.

Theorem 5.1. *If the functions* $\varphi(x), \psi(x), \phi(x), \gamma(x) \in C^4([-1,0) \cup (0,1])$, $\varphi'(\pm 1) = 0$, $\psi'(\pm 1) = 0$, $\phi'(\pm 1) = 0$, $\gamma'(\pm 1) = 0$, $\varphi(-0) = \varphi(+0)$, $\psi(-0) = \psi(+0)$, $\phi(-0) = \phi(+0)$, $\gamma(+0) = \gamma(-0)$, $\varphi'(-0) = -\varphi'(+0)$, $\psi'(-0) = -\psi'(+0)$, $\gamma'(+0) = -\gamma'(-0)$, $\varphi'''(\pm 1) = \psi'''(\pm 1)$, $\phi''(-0) = -\phi''(+0)$, $\varphi''(-0) = -\varphi''(+0)$, $\gamma''(-0) = -\gamma''(+0)$, $\psi''(-0) = -\psi''(+0)$, $\varphi'''(-0) = \varphi'''(+0)$, $\phi'''(-0) = \phi'''(+0)$, $\gamma'''(-0) = \gamma'''(+0)$, $\psi'''(-0) = \psi'''(+0)$, *then there exists a unique classical solution* $u(x,t), v(x,t) \in C_{x,t}^{2,1}(\Omega_- \cup \Omega_+)$, $f(x), g(x) \in C^1([-1,0) \cup (0,1])$ *of the problem (2.1) - (2.5).*

Proof: **1. Uniqueness of the solution.**

Let functions $(u_1(x,t), v_1(x,t), f_1(x), g_1(x))$ and $(u_2(x,t), v_2(x,t), f_2(x), g_2(x))$ be solutions of the problem (2.1) - (2.5). We denote $u = u_1(x,t) - u_2(x,t)$, $f(x) = f_1(x) - f_2(x)$, $v = v_1(x,t) - v_2(x,t)$, $g(x) = g_1(x) - g_2(x)$. Then functions (u, v, f, g) satisfy the system of equations

$$\begin{cases} u_t(x,t) - \mathrm{sign}(x) u_{xx}(x,t) + a_1 u(x,t) + b_1 v(x,t) = f(x)\,, \\ v_t(x,t) - \mathrm{sign}(x) v_{xx}(x,t) + a_2 v(x,t) + b_2 u(x,t) = g(x) \end{cases} \tag{5.1}$$

with the following homogeneous conditions:

$$u_x(\pm 1, t) = 0,\ v_x(\pm 1, t) = 0, \quad 0 \le t \le T, \tag{5.2}$$

$$u(x,0) = 0, \quad v(x,0) = 0, \quad -1 \le x \le 1\,, \tag{5.3}$$

$$u(x,T) = 0, \quad v(x,T) = 0, \quad -1 \le x \le 1\,, \tag{5.4}$$

$$\begin{aligned} &u(-0,t) = u(+0,t)\,,\ u_x(-0,t) = -u_x(+0,t)\,, \\ &v(-0,t) = v(+0,t)\,,\ v_x(-0,t) = -v_x(+0,t)\,, \end{aligned} \quad 0 \le t \le T. \tag{5.5}$$

In relation to this problem, we find the following:

Problem 1. Find a function $\omega_i(x,t)$ that satisfies equation

$$\omega_{it}(x,t) - \mathrm{sign}(x)\omega_{ixx}(x,t) - k_i\omega_i(x,t) = \bar{f}_i(x) \tag{5.6}$$

in the domain $\Omega = \Omega_- \cup \Omega_+$ and the following conditions

$$\omega_{ix}(\pm 1, t) = 0, \quad 0 \le t \le T, \tag{5.7}$$

$$\omega_i(x,0) = 0, \quad -1 \le x \le 1\,, \tag{5.8}$$

$$\omega_i(x,T)=0, \quad -1\le x\le 1\,, \tag{5.9}$$

$$\omega_i(-0,t)=\omega_i(+0,t), \quad \omega_{ix}(-0,t)=-\omega_{ix}(+0,t), \quad 0\le t\le T, \tag{5.10}$$

Let's introduce the notations

$$\omega_{in}^{+}=\int\limits_{-1}^{1}\omega_i(x,t)X_n^{+}\,dx \tag{5.11}$$

$$\bar{f}_{in}^{+}=\int\limits_{-1}^{1}\bar{f}_i(x)X_n^{+}\,dx \tag{5.12}$$

Differentiating $\omega_i(x,t)$in (5.11) under the integral sign once and taking into account equation (5.6), we obtain

$$\frac{d}{dt}\omega_{in}^{+}=\int\limits_{-1}^{1}sign(x)\omega_{ixx}(x,t)X_n^{+}dx+k_i\omega_{in}^{+}+\bar{f}_{in}^{+}.$$

Integrating twice by parts the integral in the last equality, taking into account the gluing conditions (5.10) and the characteristic equations (4.6), we obtain an equation that ω_n^{+} must satisfy

$$\frac{d}{dt}\omega_{in}^{+}=(\lambda_n^{+}+k_i)\omega_{in}^{+}+\bar{f}_{in}^{+}. \tag{5.13}$$

From conditions (5.8) and (5.9), we obtain that u_n^{+} must satisfy the conditions

$$\omega_{in}^{+}(0)=0, \quad \omega_{in}^{+}(T)=0 \tag{5.14}$$

The solution to equation (5.13) has the form

$$\omega_{in}^{+}=G_ie^{(\lambda_n^{+}+k_i)t}-\frac{\bar{f}_{in}^{+}}{(\lambda_n^{+}+k_i)}, \quad G_i=\text{const}.$$

Due to conditions (5.14) we obtain

$$\begin{cases} G_i-\dfrac{\bar{f}_{in}^{+}}{(\lambda_n^{+}+k_i)}=0, \\ G_ie^{(\lambda_n^{+}+k_i)T}-\dfrac{\bar{f}_{in}^{+}}{(\lambda_n^{+}+k_i)}=0. \end{cases}$$

Let's subtract the second equation of the system from the first one.

$$G_i(1-e^{(\lambda_n^{+}+k_i)T})=0$$

Let $\lambda_n^{+}\ne -k_i,\ \forall n=1,2,\ldots$, then $G_i=0$. Therefore, $\bar{f}_{in}^{+}=0$.

Then we get $\omega_{in}^{+}=0,\ \forall n=1,2,\ldots$ Likewise $\omega_{in}^{-}=0,\ \forall n=1,2,\ldots$

So $v(x,t)\equiv 0$, $v_1(x,t)\equiv v_2(x,t)$, $u(x,t)\equiv 0$, $u_1(x,t)\equiv u_2(x,t)$, $f\equiv 0$, $f_1(x)\equiv f_2(x)$, $g(x)\equiv 0$, $g_1(x)\equiv g_2(x)$ for $\forall(x,t)\in L_2((-1;0)\cup(0;1))$. Thus, the uniqueness of the solution to the problem is proven.

2. Existence of a solution

We return to functions

$$\omega_i(x,t)=\int\limits_{T}^{t}\xi_i(x,\tau)d\tau+\bar{\psi}_{i1}(x)=\bar{\psi}_{i1}(x)+\frac{C_iX_0(x)(e^{k_it}-e^{k_iT})}{k_i}+$$

$$\sum_{n=1}^{\infty}\left(A_{in}^{-}X_n^{-}(x)\frac{e^{(\lambda_n^{-}+k_i)t}-e^{(\lambda_n^{-}+k_i)T}}{\lambda_n^{-}+k_i}-B_{in}^{+}X_n^{+}(x)\frac{1-e^{-(\lambda_n^{+}+k_i)(T-t)}}{\lambda_n^{+}+k_i}\right), \quad (x,t)\in\Omega_{-} \tag{5.15}$$

$$\omega_i(x,t)=\int\limits_0^t \xi_i(x,\tau)d\tau+\bar{\varphi}_{i2}(x)=\bar{\varphi}_{i2}(x)+\frac{C_iX_0(x)e^{k_it}}{k_i}+$$

$$\sum_{n=1}^{\infty}\left(A_{in}^-X_n^-(x)\frac{e^{(\lambda_n^-+k_i)t}-1}{\lambda_n^-+k_i}+B_{in}^+X_n^+(x)\frac{e^{-(\lambda_n^++k_i)(T-t)}-e^{-(\lambda_n^++k_i)T}}{\lambda_n^++k_i}\right),\quad (x,t)\in\Omega_+. \tag{5.16}$$

For the coefficients A_{in}^- from (4.10) we obtain

$$|A_{in}^-|\le D_i\left(\left|\bar{\varphi}_{in}^{-(2)}\right|+\left|\bar{\psi}_{in}^{-(2)}\right|+k_i(|\bar{\varphi}_{in}^-|+|\bar{\psi}_{in}^-|)\right), \tag{5.17}$$

where $D_i=\frac{1}{\left|1-e^{(\lambda_1^-+k_i)T}\right|}$. Since the functions $\varphi''(x)$ and $\psi''(x)$ are continuous on the segment $[-1,0)\cup(0,1]$, then, as is known from the theory of Fourier series, by virtue of Bessel's inequality the following series in $L_2(-1,0)\cup(0,1)$ converge:

$$\sum_{n=1}^{\infty}\left(\bar{\varphi}_{in}^{-(2)}\right)^2\le\left\|\bar{\varphi''}_i\right\|^2, \tag{5.18}$$

$$\sum_{n=1}^{\infty}\left(\bar{\psi}_{in}^{-(2)}\right)^2\le\left\|\bar{\psi''}_i\right\|^2, \tag{5.19}$$

$$\sum_{n=1}^{\infty}\left(\bar{\varphi}_{in}^{-}\right)^2\le\left\|\bar{\varphi}_i\right\|^2, \tag{5.20}$$

$$\sum_{n=1}^{\infty}\left(\bar{\psi}_{in}^{-}\right)^2\le\left\|\bar{\psi}_i\right\|^2. \tag{5.21}$$

From (5.18)-(5.21) it follows that $\left|A_{in}^-\right|$ is bounded. Similarly, the coefficients $\left|B_{in}^+\right|$ and $|C_i|$ are also bounded.

Let us estimate the moduli of the common terms from the series (5.15) and (5.16). In the region Ω_+

$$\left|C_ie^{k_it}X_0(x)\frac{1}{k_i}\right|\le\frac{1}{\sqrt{2}\,|k_i|}\,|C_i|\,e^{|k_i|T},$$

$$\left|A_{in}^-\left(\frac{e^{(\lambda_n^-+k_i)t}-1}{\lambda_n^-+k_i}\right)X_n^-(x)\right|\le\left|A_{in}^-\right|\frac{e^{|k_i|T}}{|k_i+\lambda_n^-|\cos\sqrt{\lambda_n^+}},$$

$$\left|B_{in}^+\left(\frac{e^{-(\lambda_n^++k_i)(T-t)}-e^{-(\lambda_n^++k_i)T}}{\lambda_n^++k_i}\right)X_n^+(x)\right|\le\frac{\left|B_{in}^+\right|e^{|k_i|T}}{|\lambda_n^++k_i|}.$$

In the region Ω_-

$$\left|\frac{C_iX_0(x)(e^{k_it}-e^{k_iT})}{k_i}\right|\le\frac{1}{\sqrt{2}\,|k_i|}\,|C_i|\,e^{|k_i|T},$$

$$\left|A_{in}^-\left(\frac{(e^{(\lambda_n^-+k_i)t}-e^{(\lambda_n^-+k_i)T}}{\lambda_n^-+k_i}\right)X_n^-(x)\right|\le\left|A_{in}^-\right|\frac{e^{|k_i|T}}{|\lambda_n^-+k_i|},$$

$$\left|B_{in}^+\left(\frac{1-e^{-(\lambda_n^++k_i)(T-t)}}{\lambda_n^++k_i}\right)X_n^+(x)\right|\le\left|B_{in}^+\right|\frac{e^{|k_i|T}}{|\lambda_n^++k_i|\cos\sqrt{\lambda_n^+}}.$$

Taking into account that $\lambda_n^\pm\ne k_i$, we obtain $\frac{1}{\left|\lambda_n^\pm+k_i\right|}\le\frac{1}{\left|\lambda_n^\pm\right|q}$, where $q=\min\limits_n\left(\left|1+\frac{k_i}{\lambda_n^\pm}\right|\right)$.

If $n \to \infty$, then $|\lambda_n^{\pm}| \to \infty$ and $\frac{\sqrt{2}}{2} < \cos\sqrt{|\lambda_n^{\pm}|} < \frac{\sqrt{3}}{2}$ (see [2]). For series (5.15) and (5.16) it is true the following

$$\frac{P_i}{q}\sum_{n=1}^{\infty}\frac{\left|A_{in}^{-}\right|+\left|B_{in}^{+}\right|}{|\lambda_n^{\pm}|} \leq \frac{P_i}{q}C\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{P_i}{q}C\frac{\pi^2}{6},$$

where $P_i = 2e^{|k_i|T}$.

Therefore, the series converges absolutely and uniformly in a closed domain Ω by virtue of the Weierstrass criterion. Consequently, the functions $u(x,t)$, $v(x,t)$ is continuous as the assumption of a uniformly convergent series and a continuous function in the domain Ω.

Similar estimates show that the series (5.15) and (5.16) can be differentiated term by term with respect to the variable x twice, and with respect to the variable t once. For this, the absolute and uniform convergence of the series obtained by formal differentiation with respect to is proved.

Let's check the conditions of the final redefinition.

Let's consider the domain Ω_+

$$\begin{aligned}\omega_i(x,T) = \bar{\varphi}_2(x) + \frac{C_i X_0(x) e^{k_i T}}{k_i} + \sum_{n=1}^{\infty}\left(A_{in}^{-}X_n^{-}(x)\left(\frac{e^{(\lambda_n^{-}+k_i)T}-1}{\lambda_n^{-}+k_i}\right)+\right.\\ \left.B_{in}^{+}X_n^{+}(x)\left(\frac{-e^{-(\lambda_n^{+}+k_i)T}}{\lambda_n^{+}+k_i}\right)\right) = \varphi_2(x) + \frac{\bar{\varphi}_{i0}^{(2)} - \bar{\psi}_{i0}^{(2)} + \overline{k}_i(\bar{\varphi}_{i0}-\bar{\psi}_{i0})}{1-e^{k_i T}}\cdot\frac{X_0(x)e^{k_i T}}{k_i}\\ + \sum_{n=1}^{\infty}X_n^{-}(x)\left(\frac{-\bar{\varphi}_{in}^{-(2)} + \bar{\psi}_{in}^{-(2)} - k_i(\bar{\varphi}_{in}^{-} - \bar{\psi}_{in}^{-})}{\lambda_n^{-}+k_i}\right)+\\ \sum_{n=1}^{\infty}X_n^{+}(x)\left(\frac{\bar{\varphi}_{in}^{+(2)} - \bar{\psi}_{in}^{+(2)} + k_i(\bar{\varphi}_{in}^{+} - \bar{\psi}_{in}^{+})}{\lambda_n^{+}+k_i}\right)\left(\frac{-e^{-(\lambda_n^{+}+k_i)T}}{e^{(\lambda_n^{+}+k_i)T}-1}\right).\end{aligned} \tag{5.22}$$

Integrating the integrals for $\bar{\varphi}_{in}^{\pm(2)}$, $\bar{\psi}_{in}^{\pm(2)}$, $\bar{\varphi}_{i0}^{(2)}$, and $\bar{\psi}_{i0}^{(2)}$ by parts twice, we have $\bar{\varphi}_{in}^{\pm(2)} = \lambda_n^{\pm}\bar{\varphi}_{in}$, $\bar{\psi}_{in}^{\pm(2)} = \lambda_n^{\pm}\bar{\psi}_{in}$, $\bar{\varphi}_{i0}^{(2)} = 0$, and $\bar{\psi}_{i0}^{(2)} = 0$, where $\bar{\varphi}_{in}^{\pm} = (\mathrm{sign}(x)\cdot\bar{\varphi}_i(x), X_n^{\pm}(x))$, $\bar{\psi}_{in}^{\pm} = (\mathrm{sign}(x)\cdot\bar{\psi}_i(x), X_n^{\pm}(x))$, $\bar{\varphi}_{i0} = (\mathrm{sign}(x)\cdot\bar{\varphi}_i(x), X_0(x))$, $\bar{\psi}_{i0} = (\mathrm{sign}(x)\cdot\bar{\psi}_i(x), X_0(x))$.

Substituting the obtained ratios of coefficients into (5.22), we obtain the equalities $\omega_1(x,T) = \bar{\psi}_{12}(x)$, $\omega_2(x,T) = \bar{\psi}_{22}(x)$.

Similarly, we check the validity of the equalities $\omega_1(x,0) = \bar{\varphi}_{11}(x)$, $\omega_2(x,0) = \bar{\varphi}_{21}(x)$ in the region Ω_-.

Therefore, the obtained solution satisfies all the conditions of the problem.

The existence and uniqueness theorem is proved. □

References

[1] Gevrey M., Sur les equations aux derivees partielles du type parabolique. J. de Math. (1914) Vol. 10, Iss. 6, P. 105–148.

[2] Kaliev I. A., Mugafarov M. F., Fattahova O. V., Inverse problem for forward backward parabolic equation with generalized conjugation conditions. Ufa Mathematical Journal, (2011) Vol. 3, Iss. 2, P. 33–41.

[3] Levine H. A., Logarithmic convexity, first order differential inequalities and some applications. Trance. of AMS. (1970) Vol. 152, P. 299–320.

[4] Fayazov K. S., Khudayberganov Y. K., Nonlocal boundary value problem for a nonhomogeneous parabolic type equation with two degenerate lines. Uzbek Mathematical Journal, (2024) Vol. 68, Iss. 3, P. 53–65.

[5] Fayazov K. S., Khudayberganov Y. K., Pyatkov S., Conditional well-posedness of the initial-boundary value problem for a system of inhomogeneous mixed type equations with two degeneration lines. Journal of Mathematical Sciences (United States), (2023) Vol. 274, Iss. 2, P. 201–214.

[6] Khajiev I. O., Khudayberganov Y. K., Conditional stability and regularizations of an initial-boundary value problem for a system of parabolic type equations with two degeneration lines. AIP Conference Proceedings, (2024) Vol. 3119, Iss. 1, P. 030004.

[7] Fayazov K. S., Khajiev I. O., Estimation of conditional stability of the boundary-value problem for the system of parabolic equations with changing direction of time. Reports on Mathematical Physics, (2021) Vol. 88, Iss. 3, P. 419–431.

[8] Kozhanov A. I., Nonlinear loaded equations and inverse problems. Journal of Computational Mathematics and Mathematical Physics, (2004) Vol. 44, Iss. 4, P. 722–744.

[9] Bubnov B. A., Existence and uniqueness of solutions of inverse problems for parabolic and elliptic equations. Non-classical equations of mathematical physics, (1986), P. 25–29.

[10] Beals R., On an equation of mixed type from electron scattering. J. Math. Anal. Appl. (1977) Vol. 568, Iss. 1, P. 32–45.

[11] Beals R., Partial-range completeness and existence of solutions to two-way diffusion equation. J. Math. Phys. (1983) Vol. 32, Iss. 3, P. 565–584.

[12] Van Kampen N. G., On the theory of stationary waves in plasmas. Physica, (1977) Vol. 221, P. 458–472.

[13] Larkin A., Novikov V. A., Yanenko N. N., Nonlinear equations of variable type. Nauka, (1983).

[14] Monakhov V. N., Counter currents in the boundary layer. Dynamics of a continuous medium, (1998) Vol. 113, P. 107–113.

[15] Tersenov S. A., Parabolic equations with changing direction of time. Nauka, (1985).

[16] Pagani C., On the parabolic equation $\operatorname{sgn} x\,|x|^p u_y - u_{xx} = 0$ and a related one. Annali di Matematica Pura ed Applicata, (1974) Vol. 99, P. 333–399.

[17] Akhromeeva T. S., Kurdyumov S. P., Malinetskiy G. G., Samarskiy A. A., Nonstationary structures and diffusion chaos. Nauka, (1992).

[18] Beals R., Partial-range completeness and existence of solutions to two-way diffusion equation. J. Math. Phys. (1983) Vol. 32, Iss. 3, P. 565–584.

Fayazov K.S.,
Department of Natural and Mathematical Science, Turin Polytechnic University in Tashkent, Tashkent, Uzbekistan
e-mail: kudratillo52@mail.ru

Juraeva D.Sh.,
Department of Natural and Mathematical Science, Turin Polytechnic University in Tashkent, Tashkent, Uzbekistan
e-mail: jurayevadildora1998@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 68–73
DOI: 10.29229/uzmj.2026-2-8

Absorbing quadratic stochastic operators in genetics

Ganikhodjaev N.

Abstract. In this paper, we present exact solution of haploid Markov chain model that describes the fluctuation of gene frequency under the influence of mutation. Also generalizing Markov chain model we introduce and investigate a model that describes hybridization's process by quadratic stochastic operator.

Keywords: Markov chain, quadratic stochastic operator, gene, allele, hybrid
MSC (2020): 37N25, 60J10, 92D25

1. Genetic Motivation

Sewall Wright pointed out in an article written in 1949 [1] that the elementary evolutionary process in a reasonably large homogeneous population maybe considered to be change in gene frequency. This idea is central to population genetics, which studies the genetic makeup of populations and how it changes over time.

Recall the necessary notions from population genetics. A vague, but nevertheless informative, definition of a gene is simply a unit of hereditary information. The genetic code of an organism is carried on chromosomes. Each gene on a chromosome has different forms that it can take. These forms are called alleles. Haploid cells (or organisms) carry a single set of chromosomes and diploid organisms carry a double set of chromosomes one from each parent. The genotypes are gene composition and phenotypes are gene expression. Below we will consider haploid cells with a fixed population size of $2N$ genes composed of type-a and type-A individuals. Also we have made no mention of the dominant or recessive properties of the alleles.

2. Absorbing Markov chains

An absorbing Markov chain is a type of Markov chain where at least one state is absorbing, meaning once entered, it cannot be left. All other states are considered transient, meaning it's possible to transition to an absorbing state from them. In essence, the system will eventually reach and stay in one of the absorbing states.

The following idealized genetics model was introduced by S. Wright [2] to investigate the fluctuation of gene frequency under the influence of mutation and selection. Firstly we consider so-called simple haploid model of random reproduction, disregarding mutation pressures. Assume that we are dealing with a fixed population size of $2N$ genes composed of type-a and type-A individuals. The makeup of the next generation is determined by $2N$ independent Bernoulli trials as follows: if the parent population consists of j a-genes and $(2N-j)$ A-genes, then each trial results in a or A with probabilities

$$p_j = \frac{j}{2N}, q_j = 1 - \frac{j}{2N}$$

respectively.

Repeated selections are done with replacement. By this procedure one can generate a Markov chain X_n, where X_n is the number of a-genes in the nth generation among a constant population size of $2N$ individuals. The state space contains the $(2N+1)$ values $\{0, 1, 2, ..., 2N\}$. The transition probability matrix is computed according to the binomial distribution as

$$\Pr\{X_{n+1} = k | X_n = j\} = P_{jk} = \binom{2N}{k} p_j^k q_j^{2N-k} \tag{2.1}$$

where $j, k = 0, 1, \cdots, 2N$. Since $P_{00} = 1$ and $P_{2N,2N} = 1$, the states 0 and $2N$ are absorbing states, one of the questions of interest is to determine the probability, under the condition $X_0 = i$, that the

population will attain fixation, i.e., that it will be come a pure population composed only of a-genes or A-genes.

In [3] using first step analysis presented exact solution for absorbing Markov chains with 1 and 2 transient states.

Below we will discuss exact solution for Markov chain (2.1) with $2N = 4$. In this case we have the following transition matrix

$$\Pi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{81}{256} & \frac{27}{64} & \frac{54}{256} & \frac{3}{64} & \frac{1}{256} \\ \frac{1}{16} & \frac{4}{16} & \frac{6}{16} & \frac{4}{16} & \frac{1}{16} \\ \frac{1}{256} & \frac{3}{64} & \frac{54}{256} & \frac{27}{64} & \frac{81}{256} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Absorption occurs in states 0 and 4, and states 1,2 and 3 are transient. The probability of ultimate absorption in state 0, say, now depends on the transient state in which the process began. The first step analysis [3] proceeds by analysing, or breaking down, the possibilities that can arise at the end of the first transition, and then invoking the law of total probability coupled with the Markov property to establish a characterizing relationship among the unknown variables. Let

$$T = \min\{n \geq 0 : X_n = 0 \text{ or } X_n = 4\}$$

be the time of absorption of the process. Assume $u_i = \Pr\{X_T = 0 | X_0 = i\}$ and $\nu_i = E[T|X_0 = i]$ for $i = 1, 2, 3$.

One can extend the definitions for u_i and ν_i in a consistent and common sense manner by prescribing $u_0 = 1, u_4 = 0$, and $\nu_0 = \nu_4 = 0$. We consider separately the five contingencies $X_1 = i$ with respective probabilities P_{0i} where $i = 0, 1, 2, 3, 4$. Let

$$\begin{aligned} \Pr\{X_T = 0 | X_0 = 0\} &= 1 \\ \Pr\{X_T = 0 | X_0 = 1\} &= u_1 \\ \Pr\{X_T = 0 | X_0 = 2\} &= u_2 \\ \Pr\{X_T = 0 | X_0 = 3\} &= u_3 \\ \Pr\{X_T = 0 | X_0 = 4\} &= 0 \end{aligned}$$

We consider the three possible starting states $X_0 = 1, X_0 = 2$ and $X_0 = 3$ separately. For $X_0 = 1$ applying a first step analysis to $u_1 = \Pr\{X_T = 0 | X_0 = 1)$, the law of total probability gives

$$\begin{aligned} u_1 &= \Pr\{X_T = 0 | X_0 = 1\} \\ &= \sum_{k=0}^{4} \Pr\{X_T = 0 | X_0 = 1, X_1 = k\} \Pr\{X_1 = k | X_0 = 1\} \\ &= \sum_{k=0}^{4} \Pr\{X_T = 0 | X_1 = k\} \Pr\{X_1 = k | X_0 = 1\} \\ &= P_{10} + P_{11}u_1 + P_{12}u_2 + P_{13}u_3 \end{aligned}$$

Similarly one can produce equations for the cases $X_0 = 2$ and $X_0 = 3$. Finally we have the following system of equations

$$\begin{aligned} u_1 &= P_{10} + P_{11}u_1 + P_{12}u_2 + P_{13}u_3 \\ u_2 &= P_{20} + P_{21}u_1 + P_{22}u_2 + P_{23}u_3 \\ u_3 &= P_{30} + P_{31}u_1 + P_{32}u_2 + P_{33}u_3 \end{aligned}$$

For considered transition matrix Π we have the following system of equations

$$\begin{aligned} 148u_1 &= 81 + 54u_2 + 12u_3 \\ 10u_2 &= 1 + 4u_1 + 4u_3 \\ 148u_3 &= 1 + 12u_1 + 54u_2 \end{aligned}$$

The three equations in u_1, u_2 and u_3, are now solved simultaneously. By simple calculations we have $u_1 = 3/4; u_2 = 1/2; u_3 = 1/4$.
The result $u_3 = 1/4$ means that once begun in state $X_0 = 3$, the Markov chain ultimately end up in state 0 with probability $u_3 = 1/4$, and alternatively, will be absorbed in state 4 with probability $1 - u_3 = 3/4$.

The mean time to absorption also depends on the starting state. The first step analysis equations for $\nu_i = E[T|X_0 = i]$ are

$$\begin{aligned}\nu_1 &= 1 + P_{11}\nu_1 + P_{12}\nu_2 + P_{13}\nu_3\\ \nu_2 &= 1 + P_{21}\nu_1 + P_{22}\nu_2 + P_{23}\nu_3\\ \nu_3 &= 1 + P_{31}\nu_1 + P_{32}\nu_2 + P_{33}\nu_3\end{aligned}$$

For considered transition matrix, the equations are

$$\begin{aligned}148\nu_1 &= 256 + 54\nu_2 + 12\nu_3\\ 10\nu_2 &= 16 + 4\nu_1 + 4\nu_3\\ 148\nu_3 &= 256 + 12\nu_1 + 54\nu_2\end{aligned}$$

with solution $\nu_1 = 859/232; \nu_2 = 132/29; \nu_3 = 893/232$. Thus we have proved the following statement.

Theorem 2.1. *For $2N = 4$:*
i) a process that begins in state $X_0 = 3, (X_0 = 2$ or $X_0 = 1)$ the Markov chain ultimately end up in state 0 with probability $u_3 = 1/4$(respectively $u_2 = 1/2$ or $u_1 = 3/4$) and alternatively, will be absorbed in state 4 with probability $1 - u_i$ for $i = 1, 2, 3$;
ii) a process that begins in state $X_0 = 3, (X_0 = 2$ or $X_0 = 1)$ on the average $\nu_3 = 893/232 \approx 3,85$(respectively $\nu_2 = 132/29 \approx 4.55$ or $\nu_1 = 859/232 = 3,7$ steps will transpire prior to absorption.

2.1. **Model with mutation pressure.** A more complete model takes account of mutation pressures. We assume that prior to the formation of the new generation each gene has the possibility to mutate, that is, to change into a gene of the other kind. Specifically, we assume that for each gene the mutation $a \to A$ occurs with probability α, and $A \to a$ occurs with probability β. Again we assume that the composition of the next generation is determined by 2N independent binomial trials. The relevant values of p_j and q_j when the parent population consists of j a-genes are now taken to be

$$p_j = \frac{j}{2N}(1-\alpha) + \left(1 - \frac{j}{2N}\right)\beta \tag{2.2}$$

$$q_j = \frac{j}{2N}\alpha + \left(1 - \frac{j}{2N}\right)(1-\beta) \tag{2.3}$$

Assume that the mutation pressures operate first, after which a new gene is chosen by random selection from the population. Now, the probability of selecting an a-gene after the mutation forces have acted is just $1/(2N)$ times the number of a-genes present; hence the average probability (averaged with respect to the possible mutations) is simply $1/(2N)$ times the average number of a-genes after mutation. But this average number is clearly $j(1-\alpha) + (2N-j)\beta$, which leads at once to (2.2-2.3). The transition probabilities of the associated Markov chain are calculated by (2.1) using the values of p_j (2.2) and q_j (2.3).
If $\alpha\beta > 0$, then $0 < p_j < 1$ for all j, that is fixation will not occur in any state. Instead, as $n \to \infty$, the distribution function of X_n will approach a steady-state distribution of a random variable ξ, where $\Pr\{\xi = k\} = \pi_k$ for $k = 0, \cdots, 2N$ with $\sum_{k=0}^{2N} \pi_k = 1$. The distribution function of ξ is called the steady state gene frequency distribution.

For brevity we consider the case $2N = 4$. Then we have the following transition matrix

$$\Pi = \begin{pmatrix} (1-\beta)^4 & 4\beta(1-\beta)^3 & 6\beta^2(1-\beta)^2 & 4\beta^3(1-\beta) & \beta^4 \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} \\ \alpha^4 & 4\alpha^3(1-\alpha) & 6\alpha^2(1-\alpha)^2 & 4\alpha(1-\alpha)^3 & (1-\alpha)^4 \end{pmatrix}$$

where all entries of second, third and fourth rows are strictly positive. If $\alpha\beta > 0$ then Π is regular matrix,i.e. all entries strictly positive, therefore there is a single limit distribution, that is this matrix describes the steady-state distribution. Thus we have proved the following statement

Theorem 2.2. *For model with mutation pressure if $\alpha\beta > 0$ then one can reach the steady-state distribution.*

For $\alpha\beta = 0$ with $\alpha + \beta > 0$ absorption occurs only in state 0 or 4.

3. Hybridization

A hybrid of two genotypes refers to an individual resulting from the cross-breeding of two organisms that have different genetic makeups (genotypes). This typically involves one parent being heterozygous (carrying different alleles for a trait). The resulting offspring inherits a mix of genetic material from both parents.

Hybridization is frequently used in plant breeding to create varieties with desirable characteristics like increased yield or pest resistance. As before we assume that we are dealing with a fixed population size of $2N$ genes composed of type-a and type-A individuals. We consider a hybrid of two genotypes $X_n = i$ and $\tilde{X}_n = j$. Then one can consider genotype of size $4N$ that consists of $i+j$ a-genes and $(4N-(i+j))$ A-genes. In this case the make up of the next generation is determined by $2N$ independent Bernoulli trials as follows: If the hybrid parent population consists of $i + j$ a-genes and $(4N - (i + j))$ A-genes, then each trial results in a or A with probabilities

$$p_{ij} = \frac{i+j}{4N}, q_{ij} = 1 - \frac{i+j}{4N}$$

respectively.

Repeated selections are done with replacement. Thus the transition probability that the next generation $X_{n+1} = k$ be the result of hybrid of two genotypes $X_n = i$ and $\tilde{X}_n = j$ is computed according to the binomial distribution as

$$\Pr\{X_{n+1} = k | X_n = i, \tilde{X}_n = j\} = P_{ij,k} = \binom{2N}{k} p_{ij}^k q_{ij}^{2N-k} \tag{3.1}$$

where $i, j, k = 0, 1, \cdots, 2N$.

By cubic matrix $(P_{ij,k})_{i,j,k=0}^{2N}$ we can specify a quadratic stochastic operator $V : S^{2N} \to S^{2N}$ that is defined as follows

$$(V\mathbf{x})_k = \sum_{i,j=0}^{2N} P_{ij,k} x_i x_j, \qquad k = 0, 1, 2, \cdots, 2N, \tag{3.2}$$

where

$$S^{2N} = \{\mathbf{x} = (x_0, x_1, x_2, \cdots, x_{2N}) \in R^{2N+1} : \forall i \; x_i \geq 0, \sum_{i=0}^{2N} x_i = 1\}.$$

Such operators frequently arise in many models of mathematical genetics, namely, theory of heredity [4], [5], [6], [7],[8]. Since $p_{00} = 1$ and $p_{(2N)(2N)} = 1$, then $P_{00,0} = 1$ and $P_{(2N)(2N),2N} = 1$, the states 0 and $2N$ are absorbing states, one of the questions of interest is to determine the probability, under the condition $X_0 = i$, that the population will attain fixation, i.e., that it will be come a pure population composed only of a-genes or A-genes.

Definition 3.1. A quadratic stochastic operator with few absorbing states is called absorbing qso.

A qso (3.2) is absorbing. Let us consider example of absorbing qso in the case $2N = 4$. Then we have the following qso $V : S^4 \to S^4$

$$x_0' = x_0^2 + \frac{81}{256}x_1^2 + \frac{1}{16}x_2^2 + \frac{1}{256}x_3^2 \tag{3.3}$$
$$+\frac{2401}{2048}x_0x_1 + \frac{81}{128}x_0x_2 + \frac{625}{2048}x_0x_3 + \frac{1}{8}x_0x_4 + \frac{625}{2048}x_1x_2$$
$$+\frac{1}{8}x_1x_3 + \frac{81}{2048}x_1x_4 + \frac{81}{2048}x_2x_3 + \frac{1}{128}x_2x_4 + \frac{1}{2048}x_3x_4$$

$$x_1' = \frac{27}{64}x_1^2 + \frac{1}{4}x_2^2 + \frac{3}{64}x_3^2 \tag{3.4}$$
$$+\frac{343}{512}x_0x_1 + \frac{27}{32}x_0x_2 + \frac{375}{512}x_0x_3 + \frac{1}{2}x_0x_4 + \frac{375}{512}x_1x_2$$
$$+\frac{1}{2}x_1x_3 + \frac{135}{512}x_1x_4 + \frac{135}{512}x_2x_3 + \frac{3}{32}x_2x_4 + \frac{7}{512}x_3x_4$$

$$x_2' = \frac{27}{128}x_1^2 + \frac{3}{8}x_2^2 + \frac{27}{128}x_3^2 \tag{3.5}$$
$$+\frac{147}{1024}x_0x_1 + \frac{27}{64}x_0x_2 + \frac{675}{1024}x_0x_3 + \frac{3}{4}x_0x_4 + \frac{675}{1024}x_1x_2$$
$$+\frac{3}{4}x_1x_3 + \frac{675}{1024}x_1x_4 + \frac{675}{1024}x_2x_3 + \frac{27}{64}x_2x_4 + \frac{147}{1024}x_3x_4$$

$$x_3' = \frac{3}{64}x_1^2 + \frac{1}{4}x_2^2 + \frac{27}{64}x_3^2 \tag{3.6}$$
$$+\frac{7}{512}x_0x_1 + \frac{3}{32}x_0x_2 + \frac{135}{512}x_0x_3 + \frac{1}{2}x_0x_4 + \frac{135}{512}x_1x_2$$
$$+\frac{1}{2}x_1x_3 + \frac{375}{512}x_1x_4 + \frac{375}{512}x_2x_3 + \frac{27}{32}x_2x_4 + \frac{343}{512}x_3x_4$$

$$x_4' = \frac{1}{256}x_1^2 + \frac{1}{16}x_2^2 + \frac{81}{256}x_3^2 + x_4^2 \tag{3.7}$$
$$+\frac{1}{2048}x_0x_1 + \frac{1}{128}x_0x_2 + \frac{81}{2048}x_0x_3 + \frac{1}{8}x_0x_4 + \frac{81}{2048}x_1x_2$$
$$+\frac{1}{8}x_1x_3 + \frac{625}{2048}x_1x_4 + \frac{625}{2048}x_2x_3 + \frac{81}{128}x_2x_4 + \frac{2401}{2048}x_3x_4$$

Starting some initial point $(\tilde{x_0}, \tilde{x_1}, \tilde{x_2}, \tilde{x_3}, \tilde{x_4})$ we consider limit behaviour of corresponding trajectory. If limit exists, it will be fixed point.

For

$$S^4 = \{\mathbf{x} = (x_0, x_1, x_2, x_3, x_4) \in R^5 : \forall i \ x_i \geq 0, \sum_{i=0}^{4} x_i = 1\}$$

assume $intS^4 = \{\mathbf{x} \in S^4 : \prod_{i=0}^{4} x_i > 0\}$ is interior and $\partial S^4 = \{\mathbf{x} \in S^4 : \prod_{i=0}^{4} x_i = 0\}$ is boundary of the simplex S^4.

Let $M_0(1,0,0,0,0); M_1(0,1,0,0,0); M_2(0,0,1,0,0); M_3(0,0,0,1,0); M_4(0,0,0,0,1)$ are the vertices of the simplex.

Proposition 3.2. *For any $\boldsymbol{x} \in S^4 \setminus \{M_0, M_4\}$ we have $V(\boldsymbol{x}) \in intS^4$.*

Proof. It is evident $V(M_0) = M_0$ and $V(M_4) = M_4$. By simple calculations we have that eigenvalues of Jacobian qso V at the points M_0 and M_4 are $\lambda_1 = 2; \lambda_2 = 1; \lambda_3 = 3/8; \lambda_4 = 3/32$ and $\lambda_5 = 3/256$. Therefore M_0 and M_4 are unstable fixed points. Starting with $X_0 = 1 (X_0 = 2$, or $X_0 = 3$ we reach $V(M_1) = (81/256; 27/64; 27/128; 3/64; 1/256)$(respectively $V(M_2) = (1/16; 1/4; 3/8; 1/4; 1/16)$or $V(M_3) = (1/256; 3/64; 27/128; 27/64; 81/256)$, i.e. $V(M_1), V(M_2), V(M_3) \in intS^4$, then we can conclude validity of the proposition. □

Corollary 3.3. *For arbitrary $\boldsymbol{x} \in S^4 \setminus \{M_0, M_4\}$ we never reach pure population.*

Now we investigate the limit behaviour of the trajectories qso V.

Theorem 3.4. *Qso V is the regular transformation.*

Proof. As noted above the fixed points M_0 and M_4 are unstable fixed points. By numerical calculations one can show that $M(0.4499, 0.2634, 0.1593, 0.0895, 0.0379)$ is the unique fixed point in interior of simplex S^4. A Jacobian of qso at this fixed point is defined as follows

$$\Pi = \begin{pmatrix} 1.3415 & 0.7417 & 0.3888 & 0.1772 & 0.0679 \\ 0.3954 & 0.6950 & 0.6793 & 0.5121 & 0.3106 \\ 0.2776 & 0.3971 & 0.5579 & 0.4448 & 0.5911 \\ 0.0611 & 0.1453 & 0.2888 & 0.4679 & 0.6123 \\ 0.0144 & 0.0313 & 0.0851 & 0.2004 & 0.4181 \end{pmatrix}$$

and its eigenvalues are
$\lambda_1 = 2; \lambda_2 = 0.3786; \lambda_3 = 0.9374; \lambda_4 = 0.0705 - 0.0294i; \lambda_5 = 0.0705 + 0.0294i$.
Since all eigenvalues except $\lambda_1 = 2$ by modulo less then 1 we conclude that M is attracting fixed point, i.e V is the regular transformation. □

References

[1] Wright S., Evolution and the Genetics of Populations, Volume 2: The Theory of Gene Frequencies. Universiy of Chikago, (1969).

[2] Fisher R. A., The Genetical Theory of Natural Selection. Clarendon Press, London and New York, (1962).

[3] Taylor H. M., Karlin S., An Introduction to Stochastic Modelling. Academic Press, San Diego; London; Boston; New York; Sydney; Tokyo; Toronto, (1998).

[4] Bernstein S. N., Uchn. Zapiski. NI Kaf. Ukr. Otd. Mat. 1, 83. (1924),

[5] Jenks R. D., J. Diff. Eqs. 5, 497. (1969),

[6] Kesten H., Adv. Appl. Prob. 2, 1. (1970),

[7] Lyubich Y. I., Russian Math. Surveys 26, 51. (1978),

[8] Lyubich Y. I., Mathematical Structures in Population Genetics. Springer-Verlag, (1992).

Ganikhodjaev N.N.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan.
e-mail: nasirgani@yandex.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 74–82
DOI: 10.29229/uzmj.2026-2-9

On a mixed problem for a multidimensional elliptic equation with singular coefficients

Hasanov A., Ergashev T., Tulakova Z.

Abstract. At present the fundamental solutions of the multidimensional singular elliptic equation are known and they are expressed through the well-known Lauricella function $F_A^{(n)}$, the number of variables of which is equal to the number of singular coefficients of the equation under consideration. On the other hand, in applications of any hypergeometric function of many variables, expansion formulas are very important, allowing one to represent this hypergeometric function as an infinite sum of products of one-dimensional hypergeometric Gaussian functions for each variable of the studied function of many variables. In this paper we study one mixed problem for an elliptic equation with many singular coefficients in the first hyperoctant of the unit ball, the uniqueness of its solution is proved by the method of energy integrals, and its existence by the Green's function method. When finding the desired solution, expansion and summation formulas are used, as well as the limit relation for the Lauricella function.

Keywords: multidimensional elliptic equations with singular coefficients; fundamental solution; Gauss hypergeometric function; Lauricella function; mixed problem; Green's function.
MSC (2020): 33C05, 33C65, 35J08, 35J25, 35J70, 35J75

1. Introduction

Hypergeometric functions play a primary role in both the analysis and the solution of boundary value problems for partial differential equations, because important attributes in solving boundary value problems such as fundamental solutions and Green's functions are written in terms of hypergeometric functions. Numerous applications of hypergeometric functions can be found in mechanics, fluid mechanics, elastic dynamics, electromagnetism, and acoustics [1, 2]. The results of the theory of hypergeometric functions allow one to represent the solution of boundary value problems in explicit forms or in the form of an integral equation with a known kernel. For problems with known Green's functions expressed by hypergeometric functions, the formulation of an integral equation leads to powerful numerical approximation schemes.

Let $\mathbb{R}_m$ be the m-dimensional Euclidean space $(m \geq 2)$, $x := (x_1, ..., x_m)$ be an arbitrary point of it, and n be a positive integer such that $n \leq m$. Define the infinite region R_m^{n+}, which is a 2^{-n}th part of the Euclidean space $\mathbb{R}_m$, as follows:

$$R_m^{n+} = \{x \in \mathbb{R}_m : x_i > 0,\ i = 1, ..., n;\ -\infty < x_j < +\infty,\ j = n+1, ..., m\}.$$

In R_m^{n+}, consider the generalized singular elliptic equation

$$L_{\alpha,x}^{(m,n)}(u) \equiv \sum_{j=1}^{m} \frac{\partial^2 u}{\partial x_j^2} + \sum_{k=1}^{n} \frac{2\alpha_k}{x_k} \frac{\partial u}{\partial x_k} = 0 \tag{1.1}$$

where $\alpha := (\alpha_1, ..., \alpha_n)$ and α_k are real numbers such that $0 < 2\alpha_k < 1$ $(k = \overline{1, n})$.

Note that the fundamental solutions [3] of equation (1.1) and the particular solutions [4] of the more general equation

$$\sum_{j=1}^{k} \frac{\partial^2 u}{\partial x_j^2} + \sum_{j=1}^{k} \frac{2\alpha_j}{x_j} \frac{\partial u}{\partial x_j} = \sum_{j=k+1}^{n} \frac{\partial^2 u}{\partial x_j^2} + \sum_{j=k+1}^{n} \frac{2\alpha_j}{x_j} \frac{\partial u}{\partial x_j}, k = \overline{1, n-1}$$

are explicitly expressed in terms of the Lauricella hypergeometric function F_A^n of n variables. In this paper, using the known fundamental solution of equation (1.1), we find an unique solution of the mixed problem for equation (1.1) in the first hyperoctant of the unit ball in the explicit form.

However, only few works are devoted to the study of the mixed problem for a singular elliptic equation: in [5] was considered a two-dimensional singular elliptic equation

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}+\frac{2\alpha}{x_1}\frac{\partial u}{\partial x_1}=0,\ 0<2\alpha<1, \tag{1.2}$$

in the work [6] is discussed the mixed problem for the equation

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}+\frac{\partial^2 u}{\partial x_3^2}+\frac{2\alpha}{x_1}\frac{\partial u}{\partial x_1}=0,\ \ 0<2\alpha<1,$$

in a domain bounded in a half-space $x>0$, and the work [7] generalizes the results of the works [5, 6] to the $m-$ dimensional equation (1.1) with only one singular coefficient:

$$\sum_{i=1}^{m}\frac{\partial^2 u}{\partial x_j^2}+\frac{2\alpha}{x_1}\frac{\partial u}{\partial x_1}=0\ 0<2\alpha<1,\ m>2. \tag{1.3}$$

In the works [8] and [9], the Dirichlet and Holmgren problems were solved, respectively, by the method of potentials for equation (1.3) in the domain bounded in the subset (half-space) of the Euclidean space.

In the paper [10] for the three-dimensional elliptic equation with two singular coefficients

$$\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}+\frac{\partial^2 u}{\partial x_3^2}+\frac{2\alpha}{x_1}\frac{\partial u}{\partial x_1}+\frac{2\beta}{x_2}\frac{\partial u}{\partial x_2}=0,\ \ 0<2\alpha,2\beta<1,$$

two mixed problems and Neumann problem in the first quarter of the unit ball are studied.

The paper is organized as follows: First, in Section 2 we give some preliminary information, which will be used in what follows. Section 3 is devoted to the formulation of the problem and the uniqueness theorem. In Section 4 using the Green's functions method we find the solution of the mixed problem for equation (1.1) in the explicit form. Throughout this paper, it is assumed that the dimension of the space $m>2$.

2. Hypergeometric functions of Gauss and Lauricella

A symbol $(\kappa)_\nu$ denotes the general Pochhammer symbol or the shifted factorial, since $(1)_l=l!$ $(l\in N\cup\{0\};\ N:=\{1,2,3,...\})$, which is defined (for $\kappa,\nu\in C$), in terms of the familiar Gamma function, by

$$(\kappa)_\nu:=\frac{\Gamma(\kappa+\nu)}{\Gamma(\kappa)}=\begin{cases}1 & (\nu=0;\kappa\in C\backslash\{0\})\\ \kappa(\kappa+1)...(\kappa+l-1) & (\nu=l\in N;\kappa\in C),\end{cases}$$

it being understood conventionally that $(0)_0:=1$ and assumed tacitly that the $\Gamma-$ quotient exists.

The hypergeometric function of Gauss is defined inside the circle $|z|<1$ as the sum of the hypergeometric series [11] :

$$F(a,b;c;z)\equiv F\left[\begin{array}{l}a,b;\\ c;\end{array}\ z\right]:=\sum_{m=0}^{\infty}\frac{(a)_m(b)_m}{(c)_m\,m!}z^m,\ \ |z|<1,$$

where a, b, c, are independent of z. We shall call a, b, c the parameters of the hypergeometric function; they are arbitrary complex numbers.

If $Rec>Reb>0$, we have Euler's formula [11]

$$F(a,b;c;z)=\frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1\frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}dt. \tag{2.1}$$

Here the right-hand side is a one-valued analytic function of z within the domain $|arg(1-z)|<\pi$; therefore (2.1) gives also the analytic continuation of $F(a,b;c;z)$.

The integral representation (2.1) allows to derive the Boltz formula [11]

$$F(a,b;c;z)=(1-z)^{-b}F\left(c-a,b;c;\frac{z}{z-1}\right) \tag{2.2}$$

and get the value of the Gaussian function in unity (the summation formula) [11]

$$F(a,b;c;1)=\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)},\ Re(c-a-b)>0,\ c\neq 0,\ -1,\ -2,\ ... \tag{2.3}$$

The complete Appell hypergeometric function F_2 is defined by the series

$$F_2\left(a,b,b';c,c';x,y\right)=\sum_{m,n=0}^{\infty}\frac{(a)_{m+n}(b)_m(b')_n}{(c)_m(c')_n}\frac{x^m}{m!}\frac{y^n}{n!},\ \ |x|+|y|<1. \tag{2.4}$$

Introduce the following notation:

$$\mathbf{a}:=(a_1,...,a_n)\,,\ \ \mathbf{b}:=(b_1,...,b_n)\,,\ \ \mathbf{c}:=(c_1,...,c_n)\,;\ \ \mathbf{x}:=(x_1,...,x_n)\,.$$

In 1893 Lauricella generalized the four known Appell functions and defined his four hypergeometric functions of several variables, the first of which has the form [12]:

$$F_A^{(n)}\left(a,\mathbf{b};\mathbf{c};\mathbf{x}\right)=\sum_{k_1,k_2,...,k_n=0}^{\infty}(a)_{k_1+k_2+...+k_n}\prod_{j=1}^{n}\frac{(b_j)_{k_j}}{(c_j)_{k_j}}\frac{x_j^{k_j}}{k_j!},\quad \sum_{j=1}^{n}|x_j|<1. \tag{2.5}$$

The numerical parameters and variables of the Lauricella hypergeometric function $F_A^{(n)}$ defined in (2.5) can take complex values. It is easy to see that the function $F_A^{(n)}$ in the two-dimensional case coincides with the Appell function F_2 defined in (2.4).

The following relations hold for the Lauricella function $F_A^{(n)}$:

1) differentiation formula

$$\frac{\partial}{\partial x_k}F_A^{(n)}\left(a,\ \mathbf{b};\mathbf{c};\mathbf{x}\right)=\frac{ab_k}{c_k}F_A^{(n)}\left(a+1,\ \mathbf{b}_k+1;\mathbf{c}_k+1;\mathbf{x}\right)$$

2) the relation between contiguous hypergeometric functions [13]

$$\sum_{j=1}^{n}\frac{b_j}{c_j}x_jF_A^{(n)}\left(a+1,\mathbf{b}_j+1;\mathbf{c}_j+1;\mathbf{x}\right)=F_A^{(n)}(a+1,\mathbf{b};\mathbf{c};\mathbf{x})-F_A^{(n)}(a,\mathbf{b};\mathbf{c};\mathbf{x}),$$

where the vectors $\mathbf{b}_j+1$ and $\mathbf{c}_j+1$ are defined as follows::

$$\mathbf{b}_j+1:=(b_1,...,b_{j-1},b_j+1,b_{j+1},...,b_n)\,,$$

$$\mathbf{c}_j+1:=(c_1,...,c_{j-1},c_j+1,c_{j+1},...,c_n)\,,\ j=\overline{1,n}.$$

To study the properties of multiple hypergeometric functions, so-called expansion formulas are used, which allow one to represent a hypergeometric function of many variables through an infinite sum of products of several hypergeometric functions with one variable, and this, in turn, facilitates the process of studying the properties of functions of many variables.

Theorem 2.1. *[14, 15] For any natural $n\in \mathrm{N}\backslash\{1\}$ the following expansion formula is valid*

$$F_A^{(n)}\left(a,\mathbf{b};\mathbf{c};\mathbf{x}\right)=\sum_{|\mathbf{m}|=0}^{\infty}\frac{(a)_{A(n,n)}}{M!}\prod_{k=1}^{n}\left\{\frac{(b_k)_{B(k,n)}}{(c_k)_{B(k,n)}}x_k^{B(k,n)}F\left[\begin{matrix}a+A(k,n),b_k+B(k,n);\\ c_k+B(k,n);\end{matrix}\ x_k\right]\right\}, \tag{2.6}$$

where

$$|\mathbf{m}|=\sum_{i=1}^{n}\sum_{j=i}^{n}m_{i,j};\quad M!:=\prod_{i=1}^{n}\prod_{j=i}^{n}m_{i,j}!,\ m_{i,j}\geq 0,\ 2\leq i\leq j\leq n,$$

$$A(k,n)=\sum_{i=2}^{k+1}\sum_{j=i}^{n}m_{i,j},\ \ B(k,n)=\sum_{i=2}^{k}m_{i,k}+\sum_{i=k+1}^{n}m_{k+1,i}.$$

Theorem 2.2. *[13] Let $a, b_1, ..., b_n$ be real numbers with $a \neq 0, -1, -2, ...$ and $a > |\mathbf{b}|$. Then for $n = 1, 2, ...$ the following summation formula*

$$\sum_{|\mathbf{m}|=0}^{\infty} \frac{(a)_{A(n,n)}}{M!} \prod_{k=1}^{n} \frac{(b_k)_{B(k,n)} (a-b_k)_{A(k,n)-B(k,n)}}{(a)_{A(k,n)}} = \frac{\Gamma(a-|\mathbf{b}|)\,\Gamma^{n-1}(a)}{\prod_{k=1}^{n}\Gamma(a-b_k)} \tag{2.7}$$

is valid, where $|\mathbf{b}| := b_1 + ... + b_n$.

It is easy to see that formula (2.7) is a natural generalization of the well-known Euler formula (2.3).

Using the expansion (2.6) and the formula (2.2), it is easy to derive an analogue of the Boltz formula for the Lauricella hypergeometric function in the form

$$F_A^{(n)}(a, \mathbf{b}; \mathbf{c}; \mathbf{x}) = \prod_{k=1}^{n} (1-x_k)^{-b_k} \sum_{|\mathbf{m}|=0}^{\infty} \frac{(a)_{A(n,n)}}{M!} \prod_{k=1}^{n} \left[\frac{(b_k)_{B(k,n)}}{(c_k)_{B(k,n)}} \times \right.$$
$$\left. \times \left(\frac{x_k}{1-x_k} \right)^{B(k,n)} F\left(\begin{matrix} c_k - a + B(k,n) - A(k,n), b_k + B(k,n); \\ c_k + B(k,n); \end{matrix} \frac{x_k}{x_k - 1} \right) \right].$$

Theorem 2.3. *[16] Let a, b_k and c_k be real numbers such that $c_k \neq 0, -1, -2, ..., a > |\mathbf{b}|$ and $c_k > b_k$. Then the following limit relation is valid for $n = 1, 2, ...$:*

$$\lim_{\varepsilon \to 0} \varepsilon^{-|\mathbf{b}|} F_A^{(n)} \left[a, \mathbf{b}; \mathbf{c}; 1 - \frac{z_1(\varepsilon)}{\varepsilon}, ..., 1 - \frac{z_n(\varepsilon)}{\varepsilon} \right] = \frac{\Gamma(a-|\mathbf{b}|)}{\Gamma(a)} \prod_{k=1}^{n} \frac{\Gamma(c_k)}{[z_k(0)]^{b_k}\,\Gamma(c_k - b_k)}, \tag{2.8}$$

where $z_k(\varepsilon)$ are arbitrary functions such that $z_k(0) \neq 0$ $(k = \overline{1,n})$.

3. Formulation of the problem and the uniqueness theorem

Let $x := (x_1, ..., x_m) \in \mathbb{R}_m$ and $\Omega \subset R_m^{n+}$ is a 2^{-n} part of the m-dimensional ball with radius R and origin at the point $O(0, ..., 0)$:

$$\Omega := \{ x : \; x_1^2 + ... + x_m^2 < R^2, \; x_1 > 0, \; ..., x_n > 0, \},$$

and

$$S := \{ x : \; x_1^2 + ... + x_m^2 = R^2, \; x_1 > 0, \; ..., x_n > 0 \}$$

is the same part of the sphere corresponding to this ball.

Introduce the following notation:

$$\tilde{x}_p = (x_1, ..., x_{p-1}, x_{p+1}, ..., x_m) \in \mathbb{R}_{m-1}; \quad \mathbf{0} = (0, ..., 0) \in \mathbb{R}_{m-1}$$
$$x^{(2\alpha)} = \prod_{k=1}^{n} x_k^{2\alpha}; \; \tilde{x}_p^{(2\alpha)} = \prod_{k=1,\, k\neq p}^{n} x_k^{2\alpha_k}; \; dx = \prod_{k=1}^{m} dx_k; \; d\tilde{x}_p = \prod_{k=1,\, k\neq p}^{m} dx_p;$$
$$D_p := \left\{ \begin{matrix} x : \; x_1^2 + ... + x_{p-1}^2 + x_{p+1}^2 + ... + x_m^2 < R^2, \\ x_1 > 0, ..., x_{p-1} > 0, x_{p+1} > 0, ..., x_n > 0 \end{matrix} \right\};$$
$$S_p := \left\{ \begin{matrix} x : \; x_1^2 + ... + x_{p-1}^2 + x_{p+1}^2 + ... + x_m^2 = R^2, \\ x_1 > 0, ..., x_{p-1} > 0, x_{p+1} > 0, ..., x_n > 0 \end{matrix} \right\}; \; p = \overline{1,n}.$$

Mixed problem $\mathbf{D^nM}$. Find a regular solution $u(x) \in C(\overline{\Omega}) \cap C^2(\Omega)$ of Eq. (1.1) satisfying conditions

$$u|_{x_p=0} = \tau_p(\tilde{x}_p), \; \tilde{x}_p \in \overline{D}_p, \; p = \overline{1,n}, \quad \left. \frac{\partial u}{\partial \mathbf{N}} \right|_S = \varphi(x), \; x \in S, \tag{3.1}$$

where $\tau_p \in C(\overline{D}_p)$ and $\varphi \in C(S)$ are given smooth enough functions fulfilling the following matching conditions:

$$\tau_p(\mathbf{0}) = \tau_k(\mathbf{0}), \quad \tau_k(\tilde{x}_k)|_{x_p=0} = \tau_p(\tilde{x}_p)|_{x_k=0}, \; k \neq p \; (k, p = \overline{1,n}).$$

$\mathbf{N}$ is the outer normal to S.

Theorem 3.1. *If the problem D^nM has a solution, then it is unique.*

Proof: One can readily check the validity of the following relation:

$$x^{(2\alpha)}\left[uL_\alpha^{(m,n)}(w)-wL_\alpha^{(m,n)}(u)\right]=\sum_{j=1}^{m}\frac{\partial}{\partial x_j}\left[x^{(2\alpha)}\left(u\frac{\partial w}{\partial x_j}-w\frac{\partial u}{\partial x_j}\right)\right],\ n\le m.$$

Let Ω_ε be a sub-domain of Ω at a distance $\varepsilon>0$ from its boundary $\partial\Omega$ and $\cos(\mathbf{N},x_j)\,dD_p=d\tilde{x}_j$, if $j=p$ and $\cos(\mathbf{N},x_j)\,dD_p=0$ otherwise, where $j=\overline{1,m}$, $p=\overline{1,n}$; $\mathbf{N}$ is the outer normal to $\partial\Omega$.

Integrate both sides of this identity along the domain Ω_ε and use the Gauss-Ostrogradsky formula:

$$\int_{\Omega_\varepsilon}x^{(2\alpha)}\left[uL_\alpha^{(m,n)}(w)-wL_\alpha^{(m,n)}(u)\right]dx=\int_{\partial\Omega_\varepsilon}x^{(2\alpha)}\sum_{j=1}^{m}\left[\left(u\frac{\partial w}{\partial x_j}-w\frac{\partial u}{\partial x_j}\right)\cos(\mathbf{N},x_j)\right]d\sigma. \tag{3.2}$$

Note that Green's formula (3.2) in a more general form is proved in [17].

Using the equality

$$x^{(2\alpha)}\left[uL_\alpha^{(m,n)}(u)+\sum_{j=1}^{m}\left(\frac{\partial u}{\partial x_j}\right)^2\right]=\sum_{j=1}^{m}\frac{\partial}{\partial x_j}\left(x^{(2\alpha)}u\frac{\partial u}{\partial x_j}\right),\ n\le m,$$

we obtain

$$\int_{\Omega_\varepsilon}x^{(2\alpha)}\left[uL_\alpha^{(m,n)}(u)+\sum_{j=1}^{m}\left(\frac{\partial u}{\partial x_j}\right)^2\right]dx=\int_{\Omega_\varepsilon}\sum_{j=1}^{m}\frac{\partial}{\partial x_j}\left(x^{(2\alpha)}u\frac{\partial u}{\partial x_j}\right)dx.$$

Applying again the formula of Gauss-Ostrogradsky to this equality and letting $\varepsilon\to0$, we get

$$\int_\Omega x^{(2\alpha)}\sum_{j=1}^{m}\left(\frac{\partial u}{\partial x_j}\right)^2dx=\sum_{p=1}^{n}\int_{D_p}\tilde{x}_p^{(2\alpha)}\tau_p(\tilde{x}_p)\,f_p(\tilde{x}_p)\,d\tilde{x}_p+\int_S x^{(2\alpha)}\varphi(S)\,udS, \tag{3.3}$$

where

$$f_p(\tilde{x}_p)=\lim_{x_p\to0}x_p^{2\alpha_p}\frac{\partial u(x)}{\partial x_p},\ p=\overline{1,n}.$$

To prove the uniqueness of the solution, as usual, we suppose that the problem has two v, w solutions. Denoting $u=v-w$ we have that u satisfies homogeneous mixed problem D^nM ($\tau_p(\tilde{x}_p)=0$, $p=\overline{1,n}$, $\varphi(x)=0$). Further we have to prove that the homogeneous problem has only trivial solution. In this case from (3.3) one can easily get

$$\int_\Omega x^{(2\alpha)}\sum_{j=1}^{m}\left(\frac{\partial u}{\partial x_j}\right)^2dx=0.$$

Hence, it follows that $\dfrac{\partial u}{\partial x_j}=0$, $j=\overline{1,m}$, which implies that u is a constant function. Considering homogeneous conditions (3.1), we conclude that $u(x)\equiv0$ in $\overline{\Omega}$. The Theorem 3.1 is proved. □

4. The existence of the solution

The existence of the solution will be proved by method of Green's functions. The mixed problem D^nM on the plane ($m=2$) for an elliptic equation with one ($n=1$) singular coefficient (1.2) was solved by M.M.Smirnov [5], so here we put $m>2$.

Definition 4.1. Green's function of the mixed problem D^nM for equation (1.1) is a function $G_n(x;\xi)$, satisfying conditions:

1) this function is a regular solution of (1.1) in the domain Ω, except the point ξ;

2) it satisfies boundary conditions

$$G_n|_{x_p=0} = 0, \;\; p = \overline{1,n}, \;\; \left.\frac{\partial G_n}{\partial \mathbf{N}}\right|_S = 0,$$

3) it can be represented as

$$G_n(x;\xi) = q_n(x;\xi) + g_n(x;\xi),$$

where

$$q_n(x;\xi) = \gamma_n r^{-2\beta_n} \prod_{j=1}^{n} \left[(x_j\xi_j)^{1-2\alpha_j}\right] \cdot F_A^{(n)}\left[\begin{array}{c} \beta_n, 1-\alpha_1, ..., 1-\alpha_n; \\ 2-2\alpha_1, ..., 2-2\alpha_n; \end{array} \sigma\right]$$

is a fundamental solution [3], and function

$$g_n(x;\xi) = -\left(\frac{R}{\rho}\right)^{2\beta_n} q_n\left(x;\bar{\xi}\right),$$

is a regular solution of (1.1) in the domain Ω. Here

$$\beta_n = \frac{m-2}{2} + n - \sum_{j=1}^{n}\alpha_j, \;\; \gamma_n = 2^{2\beta_n - m}\frac{\Gamma(\beta_n)}{\pi^{m/2}}\prod_{j=1}^{n}\frac{\Gamma(1-\alpha_j)}{\Gamma(2-2\alpha_j)}, \tag{4.1}$$

$$R^2 = \sum_{j=1}^{m} x_j^2; \; \rho^2 = \sum_{j=1}^{m}\xi_j^2; \; \bar{\xi} = \left(\bar{\xi}_1, ..., \bar{\xi}_m\right), \; \bar{\xi}_j = \frac{R^2}{\rho^2}\xi_j, \;\; j = \overline{1,m}.$$

Therefore, for the domain Ω, the Green's function of the mixed problem D^nM for equation (1.1) has the form:

$$G_n(x;\xi) = q_n(x;\xi) - \left(\frac{R}{\rho}\right)^{2\beta_n} q_n\left(x;\bar{\xi}\right).$$

Next, repeating the reasoning given in [16, 18] we obtain the solution of the mixed problem in explicit form

$$u(\xi) = \gamma_n \xi^{(1-2\alpha)} \sum_{k=1}^{n}(1-2\alpha_k)\int_{S_k}\tilde{x}_k^{(1)}\tau_k(\tilde{x}_k)\left[\frac{F_A^{(n-1)}(\sigma_{k0})}{X_k^{2\beta_n}} - \frac{F_A^{(n-1)}(\bar{\sigma}_{k0})}{Y_k^{2\beta_n}}\right]dS_k + \\ + \int_S x^{(2\alpha)} G_n(x;\xi)\varphi(x)\,dS, \tag{4.2}$$

where

$$\xi^{(1-2\alpha)} = \prod_{j=1}^{n}\xi_j^{1-2\alpha_j}, \;\; \tilde{x}_k^{(1)} = \prod_{j=1, j\neq k}^{n} x_j,$$

$$F_A^{(n-1)}(z) = F_A^{(n-1)}\left[\begin{array}{c} \beta_n, 1-\alpha_1, ..., 1-\alpha_{k-1}, 1-\alpha_{k+1}, ..., 1-\alpha_n; \\ 2-2\alpha_1, ..., 2-2\alpha_{k-1}, 2-2\alpha_{k+1}, ..., 2-2\alpha_n; \end{array} z\right],$$

z is a vector with $n-1$ components and $z = \sigma_{k0}$ or $z = \bar{\sigma}_{k0}$:

$$\sigma_{k0} := \left(\sigma_1^0, ..., \sigma_{k-1}^0, \sigma_{k+1}^0, ..., \sigma_n^0\right), \quad \bar{\sigma}_{k0} := \left(\bar{\sigma}_1^0, ..., \bar{\sigma}_{k-1}^0, \bar{\sigma}_{k+1}^0, ..., \bar{\sigma}_n^0\right),$$

$$\sigma_p^0 = -\frac{4x_p\xi_p}{X_k^2}, \;\; \bar{\sigma}_p^0 = -\frac{R^2}{\rho^2}\frac{4x_p\xi_p}{Y_k^2}; \;\; X_k^2 = \xi_k^2 + \sum_{i=1, i\neq k}^{m}(\xi_i - x_i)^2,$$

$$Y_k^2 = \sum_{i=1, i\neq k}^{m}\left(R - \frac{x_i\xi_i}{R}\right)^2 + \frac{1}{R^2}\sum_{i=1, i\neq k}^{m}\sum_{j=1, j\neq i}^{m} x_i^2\xi_j^2 - (m-2)R^2.$$

Let us prove that the function defined by relation (4.2) really satisfies the mixed problem for the equation (1.1). First, we have to verify that the function

$$I_k(\xi,\sigma_{k0})=\gamma_n\xi^{(1-2\alpha)}(1-2\alpha_k)\int_{S_k}\tilde{x}_k^{(1)}\tau_k(\tilde{x}_k)X_k^{-2\beta_n}F_A^{(n-1)}(\sigma_{k0})dS_k \tag{4.3}$$

satisfies first n conditions of (3.1), i.e.

$$\lim_{\xi_l\to 0}I_k(\xi,\sigma_{k0})=\begin{cases}\tau_k\left(\tilde{\xi}_k\right), & l=k\\ 0, & l\neq k\end{cases},$$

where $\tilde{\xi}_k:=(\xi_1,...,\xi_{k-1},\xi_{k+1},...,\xi_n,...,\xi_m)\in S_k\subset\mathbb{R}_{m-1}$.

Let $l=k$ and $k\in\overline{1,n}$. Putting $x_i=\xi_i+\xi_k t_i,\ i=\overline{1,m},\ i\neq k$, we transform the formula (4.3) to the form

$$\begin{aligned}I_k(\xi,\sigma_{k0})=&(1-2\alpha_k)\gamma_n\xi^{(1-2\alpha)}\int_{T_k}\tau_k(\xi_1+\xi_k t_1,...,\xi_{k-1}+\xi_k t_{k-1},\\ &\xi_{k+1}+\xi_k t_{k+1},...,\xi_m+\xi_k t_m)\prod_{s=1,\,s\neq k}^{n}[\xi_s+\xi_k t_s]\,\tilde{R}_k^{-2\beta}F_A^{(n-1)}(\tilde{\sigma}_{k0})dT_k,\end{aligned} \tag{4.4}$$

where

$$\tilde{\sigma}_{k0}:=\left(1-\frac{\tilde{R}_{1,k}^2}{\tilde{R}_k^2},...,1-\frac{\tilde{R}_{k-1,k}^2}{\tilde{R}_k^2},1-\frac{\tilde{R}_{k+1,k}^2}{\tilde{R}_k^2},...,1-\frac{\tilde{R}_{n,k}^2}{\tilde{R}_k^2},\right)\in\mathbb{R}_{n-1},$$

$$\tilde{R}_k^2=\xi_k^2\left(1+\sum_{i=1,\,i\neq k}^{m}t_i^2\right),\ \tilde{R}_{l,k}^2=(2\xi_l+\xi_k t_l)^2+\xi_k^2\left(1+\sum_{i=1,\,i\neq l,\,i\neq k}^{m}t_i^2\right).$$

At the right-hand side of (4.4), pass to the limit as $\xi_k\to 0$. Using the limit relation (2.8) and the values of the parameter β_n and coefficient γ_n given by relation (4.1), we obtain

$$\lim_{\xi_k\to 0}I_k(\xi,\sigma_{k0})=\frac{\Gamma\left(\frac{m}{2}-\alpha_k\right)\Gamma(1-\alpha_k)}{4^{\alpha_k}\pi^{m/2}\Gamma(1-2\alpha_k)}\tau_k\left(\tilde{\xi}_k\right)\underbrace{\int_{-\infty}^{+\infty}...\int_{-\infty}^{+\infty}}_{m-1}\frac{dt_1...dt_{k-1}dt_{k+1}...dt_m}{\left(1+\sum\limits_{j=1,j\neq k}^{m}t_j^2\right)^{\frac{m}{2}-\alpha_k}}.$$

Sequentially use the identities

$$\underbrace{\int_{-\infty}^{+\infty}...\int_{-\infty}^{+\infty}}_{p}\frac{dz_1...dz_p}{\left[1+z_1^2+...+z_p^2\right]^q}=\frac{\pi^{p/2}\Gamma\left(q-\frac{p}{2}\right)}{\Gamma(q)}$$

(see [19, p. 637, Eq. 4.638(3)]), and

$$\Gamma(2z)=\frac{2^{2z-1}}{\sqrt{\pi}}\Gamma(z)\Gamma\left(z+\frac{1}{2}\right),$$

(the Legendre's duplication formula; see [11, p. 5, Eq. 1.2(15)]), we obtain

$$\lim_{\xi_k\to 0}I_k(\xi,\sigma_{k0})=\tau_k\left(\tilde{\xi}_k\right).$$

In the same way, we prove that

$$\lim_{\xi_l\to 0}I_k(\xi,\sigma_{k0})=0,\ k\neq l,\ l=1,...,n;\quad\lim_{\xi_k\to 0}I_k(\xi,\tilde{\sigma}_{k0})=0,\ 1\leq k\leq n.$$

The fulfillment of the remaining conditions of the mixed problem is established in a similar manner.

Hence, the main result of the paper is formulated as the following theorem:

Theorem 4.2. *If $\tau_p(\tilde{x}_p)\in C^2\left(\overline{D}_p\right),\ p=\overline{1,n},\ \varphi\in C^1(S)$, then the mixed problem D^nM has unique solution represented by formula (4.2).*

5. Conclusion

It's known that the potential theory has played a paramount role in both analysis and computation for boundary value problems for elliptic equations. Recently, the potential theories only for the two-dimensional elliptic equation with two singular coefficients [20] and the multidimensional elliptic equation with one singular coefficient [7] are constructed. In the study of potentials, the properties of the fundamental solutions of the given equation are essentially and fruitfully used. At the present time, fundamental solutions of the multidimensional elliptic equation with several singular coefficients are already known (see [3, 21]).

Despite the fact that fundamental solutions, particular solutions and expansion formulas are known even for a more general equation, the construction of potential theory for a singular elliptic equation was limited to the equations with one and two singular coefficients. For the future, it would be interesting to apply potential theory to the solution of boundary value problems for elliptic equations with two or more singular coefficients, for example, in the beginning for the equation

$$u_{xx} + u_{yy} + u_{zz} + \frac{2\alpha}{x}u_x + \frac{2\beta}{y}u_y + \frac{2\gamma}{z}u_z = 0,\ 0 < 2\alpha, 2\beta, 2\gamma < 1,\ x > 0, y > 0, z > 0.$$

References

[1] Bers L., Mathematical Aspects of Subsonic and Transonic Gas Dynamics. Wiley, New York, (1958).

[2] Niukkanen A. W., Generalised hypergeometric series ${}^{N}F(x_1, ..., x_N)$ arising in physical and quantum chemical applications. J. Phys. A: Math. Gen. (1983) Vol. 16, P. 1813–1825.

[3] Ergashev T. G., Fundamental solutions for a class of multidimensional elliptic equations with several singular coefficients. Journal of Siberian Federal University. Mathematics and Physics, (2020) Vol. 13, Iss. 1, P. 48–57.

[4] Ryskan A. R., Arzikulov Z. O., Ergashev T. G., Particular solutions of multidimensional generalized Euler-Poisson-Darboux equations of various (elliptic or hyperbolic) types. Journal of Mathematics, Mechanics and Computer Science, (2024) Vol. 1, Iss. 121, P. 76–88.

[5] Smirnov M. M., Boundary mixed problem for the equation $y^{\mu}u_{xx} + u_{yy} = 0$. Siberian Mathematical Journal, (1963) Vol. 4, Iss. 5, P. 1150–1161.

[6] Ergashev T. G., Potentials for three-dimensional singular elliptic equation and their application to the solving a mixed problem. Lobachevskii Journal of Mathematics, (2020) Vol. 41, Iss. 6, P. 1067–1077.

[7] Tulakova Z. R., Ergashev T. G., Application of the potential theory to solving a mixed problem for a multidimensional elliptic equation with one singular coefficient. KazNU Bulletin. Mathematics, Mechanics, Computer Science Series, (2024) Vol. 3, Iss. 123, P. 33–47.

[8] Ergashev T. G., Potentials for the singular elliptic equations and their application. Results in Applied Mathematics, (2020) Vol. 7, P. 100126.

[9] Srivastava H. M., Hasanov A., Ergashev T. G., A family of potentials for elliptic equations with one singular coefficient and their applications. Mathematical Methods in Applied Sciences, (2020) Vol. 43, Iss. 10, P. 6181–6199.

[10] Tulakova Z. R., Spatial mixed problems and Neumann problem for the three-dimensional elliptic equation with the two singular coefficients. Uzbek Mathematical Journal, (2024) Vol. 68, Iss. 3, P. 150–157.

[11] Erdélyi A., Magnus W., Oberhettinger F., Tricomi F. G., Higher Transcendental Functions, 1. McGraw-Hill, New York, Toronto and London, (1953).

[12] Lauricella G., Sulle funzioni ipergeometriche a piu variabili. Rend. Circ. Mat. Palermo, (1893) Vol. 7, P. 111–158.

[13] Ergashev T. G., Generalized Holmgren problem for an elliptic equation with several singular coefficients. Differential Equations, (2020) Vol. 56, Iss. 7, P. 842–856.

[14] Hasanov A., Ergashev T. G., New decomposition formulas associated with the Lauricella multivariable hypergeometric functions. Montes Taurus Journal of Pure and Applied Mathematics, (2021) Vol. 3, Iss. 3, P. 317–326.

[15] Ergashev T. G., Ryskan A., Yuldashev N. N., Recurrence free decomposition formulas for the Lauricella special functions. Bulletin of the Karaganda University. Mathematics Series, (2024) Vol. 4, Iss. 116, P. 95–106.

[16] Ergashev T. G., Tulakova Z. R., Dirichlet problem for an elliptic equation with several singular coefficients in an infinite domain. Russian Mathematics, (2021) Vol. 65, Iss. 7, P. 71–80.

[17] Ekincioğlu I., Guliyev V. S., Shishkina E. L., Fractional weighted spherical mean and maximal inequality for the weighted spherical mean and its application to singular PDE. Journal Mathematical Sciences, (2022) Vol. 266, P. 744–764.

[18] Ergashev T. G., Tulakova Z. R., The Neumann problem for a multidimensional elliptic equation with several singular coefficients in an infinite domain. Lobachevskii Journal of Mathematics, (2022) Vol. 43, Iss. 1, P. 199–206.

[19] Gradshteyn I. S., Ryzhik I. M., Table of Integrals, Series, and Products. Academic Press, Amsterdam, (2007).

[20] Hasanov A., Ergashev T. G., On potential theory for the generalized bi-axially symmetric elliptic equation in the plane. Journal of Mathematics, Mechanics and Computer Sciences, (2021) Vol. 1, Iss. 109, P. 3–24.

[21] Ergashev T. G., Hasanov A., Fundamental solutions of the bi-axially symmetric Helmholtz equation. Uzbek Mathematical Journal, (2018) Vol. 1, P. 55–64.

Hasanov Anvardjan,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan,
University of Ghent, Belgium
e-mail: anvarhasanov@yahoo.com

Ergashev Tuhtasin,
Tashkent Institute of Irrigation and, Agricultural Mechanization Engineers, – National Research University,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan,
University of Ghent, Belgium
e-mail: ergashev.tukhtasin@gmail.com

Tulakova Ziyodaxon,
Fergana State Technical University, Fergana, Uzbekistan
e-mail: ziyodacoders@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 83–89
DOI: 10.29229/uzmj.2026-2-10

An evasion game for an infinite system of ternary differential equations in the Hilbert space l_2

Ibragimov G., Muxammadjonov A., Qushaqov X.

Abstract. In this paper, an evasion game for an infinite system of ternary differential equations is studied. The game is considered in the Hilbert space l_2. The control parameters of pursuer and evader are subject to geometric constraints. The purpose is to construct strategies for the evader to avoid being captured in the game. Equations for guaranteed evasion time for the evader are derived, describing how long the evader can avoid capture. Furthermore, this research paper contributes to understanding the concept of a guaranteed evasion game.

Keywords: Differential game, evader, control, strategy, infinite system of differential equations, geometric constraint
MSC (2020): 49N70, 49N75, 93C15

1. Introduction

In the mid of 20th century, the theory of differential games, a fascinating intersection of control theory and game theory, were formally established by R.Isaacs [1], who first used the concept of "Differential Games" in his work. This work inspired numerous researchers including Pontryagin L.S. [2, 3], Krasovskii N.N. [4] and others who obtained the fundamental results on differential games.

In the last decades, on the other hand, the researchers such as Azamov, Blagodatskikh, Chikrii, Ibragimov, Mamadaliev, Petrov, Samatov have been obtaining interesting and important results (see [5]-[6], [7]-[8], [9]-[10], [11]-[12]). Azamov and Ruziboev [5] explored time-optimal control problems associated with PDEs, while Blagodatskikh and Petrov's investigations of group pursuit strategies and impulsive controls provided critical insights into strategies for dynamic systems ([13, 14] and [9]-[10]). However, Chikrii's work [15, 16] provided critical insights into linear differential games with integral constraints.

A great advancement in differential games involved reducing differential game problems modelled as PDE to infinite systems of differential equations, methods like Fourier analysis. We can see using such transformation in studies by Ibragimov et al. [17], [18]-[19], who addressed infinite-dimensional pursuit-evasion games in the Hilbert space l_2.

The works of Satimov and Tukhtasinov [20],[21] focused on pursuit and evasion games in systems described by parabolic type equations. In these games, the conditions that guarantees pursuit and evasion were obtained. Other studies, such as Mamadaliev et al. [7]-[8], explored differential pursuit games with integral constraints and impulse controls. In the paper by Ibragimov [22], an evasion differential game of many pursuers and one evader was considered. The evader has an advantage over pursuers in control resource. A sufficient condition of an evasion was found and an evasion strategy was constructed. Moreover, in the paper by Kuchkarov et al. [23], a simple motion differential game of many pursuers and one evader was studied on manifolds with Euclidean metric. All players have equal dynamic capabilities. The condition that ensures a pursuit can be completed in the game was obtained, if the condition is not satisfied, an evasion is possible.

The works [24, 20, 21] are devoted to a differential game described by a PDE of the form

$$\frac{\partial f}{\partial t} = Af + u - v, \quad Af = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial f}{\partial x_i}\right).$$

In these works, the given equation was reduced the equation to infinite system of differential equations

$$\dot{f}_k + \lambda_k f_k = u_k - v_k, \ k = 1, 2, ..., \tag{1.1}$$

where v_k and u_k are control parameters of evader and pursuer, respectively, $f_k, u_k, v_k \in \mathbb{R}$, and coefficients λ_k, $k = 1, 2, ...$, satisfy the condition $0 < \lambda_1 \leq \lambda_2 \leq ... \to \infty$.

In the papers [25], [26], the following evasion differential game for an infinite system of binary differential equations

$$\begin{array}{ll} \dot{x}_i = -\lambda_i x_i + y_i + u_{i1} - v_{i1}, & x_i(0) = x_{i0}, \\ \dot{y}_i = -\lambda_i y_i + u_{i2} - v_{i2}, & y_i(0) = y_{i0}, \quad i = 1, 2, ..., \end{array} \tag{1.2}$$

was studied in the Hilbert space l_2.

The present paper aims to contribute to an evolving field by investigating an evasion game for an infinite system of ternary differential equations within the Hilbert space l_2. In this game, the control parameters of the players are subject to geometric constraints. We develop a strategy for the evader that guarantees an evasion during the game. Additionally, we provide an equation to determine the guaranteed evasion time.

1.1. **Statement of problem.** It is important to highlight that the Hilbert space under consideration is l_2, which is a vector space:

$$l_2 = \left\{ \alpha = (\alpha_1, \alpha_2, ..., \alpha_n, ...) \mid \sum_{n=1}^{\infty} \alpha_n^2 < \infty \right\},$$

whose elements are sequences of real numbers.

Moreover, the inner product and norm are defined as follows:

$$(\alpha, \beta) = \sum_{n=1}^{\infty} \alpha_n \beta_n, \quad ||\alpha|| = \sqrt{(\alpha, \alpha)}.$$

We examine a controlled object given by the following infinite system of differential equations

$$\begin{array}{ll} \dot{x}_i = -\lambda_i x_i + y_i + u_{i1} - v_{i1}, & x_i(0) = x_{i0}, \\ \dot{y}_i = -\lambda_i y_i + z_i + u_{i2} - v_{i2}, & y_i(0) = y_{i0}, \\ \dot{z}_i = -\lambda_i z_i + u_{i3} - v_{i3}, & z_i(0) = z_{i0} \quad i = 1, 2, 3, ..., \end{array} \tag{1.3}$$

in the Hilbert space l_2, where λ_i are given positive numbers, $x_0 = (x_{10}, x_{20}, x_{30}, \dots) \in l_2$, $y_0 = (y_{10}, y_{20}, y_{30}, \dots) \in l_2$, $z_0 = (z_{10}, z_{20}, z_{30}, \dots) \in l_2$ are initial positions, $u = (u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23}, \dots)$ and $v = (v_{11}, v_{12}, v_{13}, v_{21}, v_{22}, v_{23}, \dots)$ are the control parameters of pursuer and evader, respectively. We suppose that $0 \leq t \leq T$, where T is a given sufficiently large number, and that $\varphi_0 = (\varphi_{10}, \varphi_{20}, \varphi_{30}, ...) = (x_{10}, y_{10}, z_{10}, x_{20}, y_{20}, z_{20}, x_{30}, y_{30}, z_{30}, \dots) \neq 0$.

Assume that ρ and σ, $\rho > \sigma$, are given positive numbers.

Definition 1.1. A function $u(t) = (u_1(t), u_2(t), u_3(t), \dots)$, $t \in [0, T]$, with measurable coordinates $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, $i = 1, 2, ...$, subject to the condition

$$\sum_{i=1}^{\infty} (u_{i1}^2(t) + u_{i2}^2(t) + u_{i3}^2(t)) \leq \rho^2, \quad 0 \leq t \leq T, \tag{1.4}$$

is referred to as the admissible control of the pursuer.

Definition 1.2. A function $v(t) = (v_1(t), v_2(t), v_3(t), \dots)$, $t \in [0, T]$, with measurable coordinates $v_i(t) = (v_{i1}(t), v_{i2}(t), v_{i3}(t))$, $i = 1, 2, ...$, subject to the condition

$$\sum_{i=1}^{\infty} (v_{i1}^2(t) + v_{i2}^2(t) + v_{i3}^2(t)) \leq \sigma^2, \quad 0 \leq t \leq T, \tag{1.5}$$

is referred to as the admissible control of the evader.

We define $S(\sigma)$ as the set of all admissible controls.

Let

$$A_i = \begin{bmatrix} -\lambda_i & 1 & 0 \\ 0 & -\lambda_i & 1 \\ 0 & 0 & -\lambda_i \end{bmatrix}, \quad i = 1, 2, ...,$$

then, we have

$$e^{A_i t} = e^{-\lambda_i t} \begin{bmatrix} 1 & t & \frac{1}{2}t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}, \quad i = 1, 2, ... \tag{1.6}$$

Let $C(0, T; l_2)$ be the space of continuous functions $\varphi(\cdot)$ such that $\varphi(t) \in l_2$ for each $0 \le t \le T$. We present the following proposition which is proved in [6].

Proposition 1.1. If $w(\cdot) \in S(\sigma)$, and $\lambda_i \ge 0$, $i = 1, 2, ...$, then there exists a unique solution $\varphi(t) = (\varphi_1(t), \varphi_2(t), ...)$ of the infinite system of differential equations (1.3) in the space $C(0, T; l_2)$.

Of course, for $i = 1, 2, ...$, we have (see [6])

$$\varphi_i(t) = e^{A_i t}\varphi_{i0} + \int_0^t e^{A_i(t-s)}(u_i(s) - v_i(s))ds, \quad i = 1, 2, ...$$

By denoting $\varphi_i(t) = e^{A_i t}\psi_i(t)$, it is not difficult to see that $\varphi(t) = 0$ implies $\psi(t) = 0$, where

$$\psi(t) = (\psi_1(t), \psi_2(t), ...), \quad \psi_i(t) = \varphi_{i0} + \int_0^t e^{-A_i s}(u_i(s) - v_i(s))ds, \quad i = 1, 2, ... \tag{1.7}$$

Let $\varphi_i(t) = (x_i(t), y_i(t), z_i(t))$, $\varphi_{i0} = (x_{i0}, y_{i0}, z_{i0})$, $v_i = (v_{i1}, v_{i2}, v_{i3})$.

Definition 1.3. We call the function

$$(t, \varphi, \xi, \eta, u) \to V(t, \varphi, \xi, \eta, u), \quad V : \mathbb{R}_+ \times l_2 \times [0, \rho] \times [0, \sigma] \times l_2 \to l_2,$$

the strategy of the evader, if, for any admissible control of pursuer $u = u(t)$, $t \in [0, T]$, the system with $v = V(t, \varphi(t), \rho(t), \sigma(t), u(t))$, has a unique solution

$$\varphi(t) = (\varphi_1(t), \varphi_2(t), \ldots) = (x_1(t), y_1(t), z_1(t), x_2(t), y_2(t), z_2(t), \ldots), \quad i = 1, 2, ...$$

after substitution of the players' controls into it.

Definition 1.4. We say that the strategy V of evader guarantees evasion on the interval $[0, \tau(V))$ from the initial state $\varphi_0 \in l_2$ if $\varphi(t) \ne 0$, $t \in [0, \tau(V))$, for any pursuer's admissible control $u(t)$, $t \in [0, T]$. We call the number $\tau(V)$ a guaranteed evasion time.

The evader is interested in maximizing the guaranteed evasion time θ by choosing its strategy V but the pursuer tries to minimize θ by using its control $u = u(t)$, $0 \le t \le T$.

Problem 1. Construct an admissible control of the evader and find an equation for a guaranteed evasion time θ in game (1.3).

2. Main result

In this section we present an admissible control for the evader and determine a guaranteed evasion time θ in game (1.3).

Theorem 2.1. For any initial state $\varphi_0 = (\varphi_{10}, \varphi_{20}, \cdots) \ne 0$, the number

$$\theta = \sup_i \theta_i,$$

is a guaranteed evasion time in game (1.2), where

$$\theta_i = \frac{1}{\lambda_i} \ln \frac{-(\rho - \sigma) + \sqrt{(\rho - \sigma)^2 + 2\lambda_i |\varphi_{i0}|(\rho + \sigma)}}{\rho + \sigma}$$

if $\lambda_i \ge 1$, and

$$\theta_i = \frac{1}{2} \ln \left(\frac{2|\varphi_{i0}|}{\rho + \sigma} + 1 \right)$$

if $0 < \lambda_i < 1$.

Proof. To prove theorem, we construct a strategy for the evader that ensures $\varphi(t) \neq 0$ on $[0, \theta']$ where θ' is an arbitrary time satisfying the condition $0 < \theta' < \theta$, while the pursuer can use an arbitrary control $u = u(t)$.

Indeed, we can find $j \in \{1, 2, \cdots\}$ such that $\theta' < \theta_j$ by definition of θ. Thus, our main goal is to show that in the game (1.3), an evasion is possible on $[0, \theta_j)$. Observe $\varphi_{j0} \neq 0$ since otherwise $\theta_j = 0$ contradicting positivity of θ_j.

We present the control for the evader as follows:

$$\begin{aligned} v_j(t) &= -\sigma \frac{1}{1 + t + \frac{1}{2}t^2} \begin{bmatrix} 1 & t & \frac{1}{2}t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} e_j, \\ v_i(t) &= 0, \ i = 1, 2, \cdots, \ i \neq j, \quad t \in [0, \theta_j). \end{aligned} \tag{2.1}$$

where $e_j = \frac{\varphi_{j0}}{|\varphi_{j0}|}$.

Lemma 2.1. Control (2.1) serves is the admissible control for the evader.

Proof. To prove the lemma, we use (1.5)

$$\begin{aligned} |v_j(t)| &= \left| -\sigma \frac{1}{1 + t + \frac{1}{2}t^2} \begin{bmatrix} 1 & t & \frac{1}{2}t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e_{j1} \\ e_{j2} \\ e_{j3} \end{bmatrix} \right| = \frac{\sigma}{1 + t + \frac{1}{2}t^2} \left| \begin{bmatrix} e_{j1} + te_{j2} + \frac{1}{2}t^2 e_{j3} \\ e_{j2} + te_{j3} \\ e_{j3} \end{bmatrix} \right| \\ &\leq \frac{\sigma}{1 + t + \frac{1}{2}t^2} \left| \begin{bmatrix} e_{j1} \\ e_{j2} \\ e_{j3} \end{bmatrix} + \begin{bmatrix} te_{j2} \\ te_{j3} \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{2}t^2 e_{j3} \\ 0 \\ 0 \end{bmatrix} \right| \leq \frac{\sigma}{1 + t + \frac{1}{2}t^2} \left(1 + t + \frac{1}{2}t^2 \right) = \sigma. \end{aligned}$$

Consequently, control (2.1) is admissible. □

For any admissible control $u(t)$ of the pursuer, where time $t \in [0, \theta_j)$, we have

$$\begin{aligned} \psi_j(t) &= \varphi_{j0} + \int_0^t e^{-A_j s} u_j(s) ds - \int_0^t e^{-A_j s} v_j(s) ds \\ &= \varphi_{j0} + \int_0^t e^{-A_j s} u_j(s) ds + \sigma e_j \int_0^t \frac{1}{1 + s + \frac{1}{2}t^2} e^{-\lambda_j s} ds. \end{aligned} \tag{2.2}$$

From the following inequality

$$|(e^{-A_j s} u_j(s), e_j)| \leq |e^{-A_j s} u_j(s)||e_j| = |(e^{-A_j s} u_j(s)| = e^{\lambda_j s} \left| \begin{bmatrix} 1 & -s & \frac{1}{2}s^2 \\ 0 & 1 & -s \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_{j1} \\ u_{j2} \\ u_{j3} \end{bmatrix} \right| \leq \rho e^{\lambda_j s} \left(1 + s + \frac{1}{2}s^2 \right)$$

we have

$$|(e^{-A_j s} u_j(s), e_j)| \leq \rho e^{\lambda_j s} \left(1 + s + \frac{1}{2}s^2 \right). \tag{2.3}$$

Now, we multiple both sides of equation (2.2) by e_j and by using (2.3), we obtain

$$\begin{aligned} \psi_j(t) e_j &= |\varphi_{j0}| + \int_0^t \left(e^{-A_j s} u_j(s), e_j \right) ds + \sigma \int_0^t \frac{1}{1 + s + \frac{1}{2}s^2} e^{\lambda_j s} ds \\ &\geq |\varphi_{j0}| - \rho \int_0^t \left(1 + s + \frac{1}{2}s^2 \right) e^{\lambda_j s} ds + \sigma \int_0^t \frac{1}{1 + s + \frac{1}{2}s^2} e^{\lambda_j s} ds \\ &= |\varphi_{j0}| - (\rho + \sigma) \int_0^t \left(1 + s + \frac{1}{2}s^2 \right) e^{\lambda_j s} ds + \\ &+ \sigma \int_0^t \left[\left(1 + s + \frac{1}{2}s^2 \right) e^{\lambda_j s} + \frac{1}{1 + s + \frac{1}{2}s^2} e^{\lambda_j s} \right] ds. \end{aligned}$$

Hence, by the arithmetic mean-geometric mean inequality we obtain

$$\begin{aligned}\psi_j(t)e_j >&|\varphi_{j0}| - (\rho+\sigma)\int_0^t (1+s+\frac{1}{2}s^2)e^{\lambda_j s}ds + 2\sigma\int_0^t e^{\lambda_j s}ds\\ =&|\varphi_{j0}| - (\rho+\sigma)\int_0^t (s+\frac{1}{2}s^2)e^{\lambda_j s}ds - (\rho+\sigma)\int_0^t e^{\lambda_j s}ds + 2\sigma\int_0^t e^{\lambda_j s}ds\\ =&|\varphi_{j0}| - (\rho+\sigma)\int_0^t (s+\frac{1}{2}s^2)e^{\lambda_j s}ds - (\rho-\sigma)\int_0^t e^{\lambda_j s}ds.\end{aligned}$$

Then, by using inequality $s+\frac{1}{2}s^2 < e^s$ we have the following

$$\psi_j(t)e_j > |\varphi_{j0}| - (\rho+\sigma)\int_0^t e^{(1+\lambda_j)s}ds - (\rho-\sigma)\int_0^t e^{\lambda_j s}ds.$$

We now consider two cases according to the values of λ_j. The first case is when $\lambda_j \geq 1$ and the second one is $0 < \lambda_j < 1$.

In the first case, by the condition $\lambda_j \geq 1$ we have

$$\begin{aligned}\psi_j(t)e_j >&|\varphi_{j0}| - (\rho+\sigma)\int_0^t e^{(1+\lambda_j)s}ds - (\rho-\sigma)\int_0^t e^{\lambda_j s}ds\\ \geq&|\varphi_{j0}| - (\rho+\sigma)\int_0^t e^{2\lambda_j s}ds - (\rho-\sigma)\int_0^t e^{\lambda_j s}ds.\end{aligned}$$

Since $t \in [0, \theta_j]$,

$$\begin{aligned}\psi_j(t)e_j >&|\varphi_{j0}| - (\rho+\sigma)\int_0^{\theta_j} e^{2\lambda_j s}ds - (\rho-\sigma)\int_0^{\theta_j} e^{\lambda_j s}ds\\ =&|\varphi_{j0}| - (\rho+\sigma)\frac{e^{2\lambda_j\theta_j}-1}{2\lambda_j} - (\rho-\sigma)\frac{e^{\lambda_j\theta_j}-1}{\lambda_j}\\ =&-(\rho+\sigma)\frac{e^{2\lambda_j\theta_j}}{2\lambda_j} + \frac{\rho+\sigma}{2\lambda_j} + |\varphi_{j0}| - (\rho-\sigma)\frac{e^{\lambda_j\theta_j}}{\lambda_j} + \frac{(\rho-\sigma)}{\lambda_j}\\ =&-(\rho+\sigma)\frac{e^{2\lambda_j\theta_j}}{2\lambda_j} - (\rho-\sigma)\frac{e^{\lambda_j\theta_j}}{\lambda_j} + |\varphi_{j0}| + \frac{(3\rho-\sigma)}{2\lambda_j}\\ >&-(\rho+\sigma)\frac{e^{2\lambda_j\theta_j}}{2\lambda_j} - (\rho-\sigma)\frac{e^{\lambda_j\theta_j}}{\lambda_j} + |\varphi_{j0}|.\end{aligned}$$

By the value of θ_j, we have

$$\psi_j(t)e_j > -(\rho+\sigma)\frac{e^{2\lambda_j\theta_j}}{2\lambda_j} - (\rho-\sigma)\frac{e^{\lambda_j\theta_j}}{\lambda_j} + |\varphi_{j0}| = 0. \tag{2.4}$$

In the second case, where $0 < \lambda_j < 1$, we obtain

$$\begin{aligned}\psi_j(t)e_j >&|\varphi_{j0}| - (\rho+\sigma)\int_0^t e^{(1+\lambda_j)s}ds + 2\sigma\int_0^t e^{\lambda_j s}ds\\ >&|\varphi_{j0}| - (\rho+\sigma)\int_0^t e^{(1+\lambda_j)s}ds > |\varphi_{j0}| - (\rho+\sigma)\int_0^t e^{2s}ds\\ >&|\varphi_{j0}| - (\rho+\sigma)\int_0^{\theta_j} e^{2s}ds = |\varphi_{j0}| - (\rho+\sigma)\frac{e^{2\theta_j}-1}{2} = 0.\end{aligned}$$

Hence, in both cases, we obtained that $\psi_j(t)e_j > 0, 0 \leq t < \theta_j$ which implies $\psi_j(t) \neq 0$. As the result, $\varphi_j(t) \neq 0$ by the equation (1.7), thus, $\varphi(t) \neq 0$ for $t \in [0, \theta_j)$. In particular, $\varphi(t) \neq 0$ on the interval $[0, \theta']$. Proof of Theorem 2.1 is complete. □

3. Conclusions

In this research paper, we have focused on investigating the guaranteed evasion time within the context of the game described by equation (1.3), while an evasion game for an infinite system of binary differential equations in the Hilbert space l_2 was studied in the papers [25], [26]. The central focus was to construct a guaranteed evasion strategy for the evader. The control parameters of both players are subject to geometric constraints. We derived explicit formulas to calculate the guaranteed evasion time. Future research may focus on broadening these results to study more complex constraints, multi-player scenarios, further advancing the applicability of these methods to real-world problems and also finding optimal evasion strategy is an open problem.

References

[1] Isaacs R., Differential Games. John Wiley & Sons, New York, NY, (1965).

[2] Pontryagin L. S., Linear differential games. SIAM Journal on Control, (1974) Vol. 12, Iss. 2,

[3] Pontryagin L. S., Mischenko Y. F., The problem of evasion in linear differential games. Differencial'nye Uravnenija, (1971) Vol. 7, Iss. 3, P. 436–445.

[4] Krasovskii N. N., Control of a Dynamical System. Nauka, Moscow, (1985).

[5] Azamov A. A., Ruziboev M. B., The time-optimal problem for evolutionary partial differential equations. Journal of Applied Mathematics and Mechanics, (2013) Vol. 77, Iss. 2, P. 220–224.

[6] Ibragimov G. I., Qo'shaqov H. S., Turgunov I., Alias I. A., The solution of an infinite system of ternary differential equations. Mathematical Modeling and Computing, (2022) Vol. 9, Iss. 4, P. 833–841.

[7] Mamadaliev N. A., The pursuit problem for linear games with integral constraints on players' controls. Russian Mathematics, (2020) Vol. 64, Iss. 3, P. 9–24.

[8] Mamadaliev N. A., Ibaydullayev T. T., Abduolimova G. M., On game problems of controlling pencil trajectories under integral constraint on the controls of players. Russian Mathematics, (2023) Vol. 67, P. 34–45.

[9] Petrov N. N., Multiple capture of the given number of evaders in the nonstationary example of Pontryagin L. S. 14th International Conference "Stability and Oscillations of Nonlinear Control Systems", Moscow, Russia, (2018), P. 1–3.

[10] Petrov N. N., To the time-dependent problem of group pursuit with phase constraints. Mathematical Game Theory and its Applications, (2010) Vol. 2, Iss. 4, P. 74–83.

[11] Samatov B., Horilov M. A., Juraev B. I., Π-strategy for a differential game of pursuit with integral constraints of a generalized type. Vestnik Udmurtskogo Universiteta. Matematika. Mekhanika. Komp'yuternye Nauki, (2023) Vol. 33, P. 293–311.

[12] Samatov B. T., Juraev B. I., Pursuit-evasion problems with a constraint for energy expenditure of velocity. Uzbek Mathematical Journal, (2022) Vol. 66, Iss. 4, P. 146–155.

[13] Blagodatskikh A. I., Petrov N. N., Conflict Interaction Between Groups of Controlled Objects. Udmurt State University Press, Izhevsk, (2009).

[14] Blagodatskikh A. I., Petrov N. N., Simultaneous multiple capture of rigidly coordinated evaders. Dynamic Games and Applications, (2019) Vol. 9, P. 594–613.

[15] Chikrii A. A., Belousov A. A., On linear differential games with integral constraints. Proceedings of the Steklov Institute of Mathematics, (2009) Vol. 269, Iss. Suppl. 1, P. S69–S80.

[16] Chikrii A. A., Machikhin I. I., Linear differential games with impulse control of players. Tr. Inst. Mat. Mekh. Ural. Otd. Ross. Akad. Nauk, (2005) Vol. 11, Iss. 1, P. 212–224.

[17] Alimov S. A., Ibragimov G. I., Time optimal control problem with integral constraint for the heat transfer process. Eurasian Mathematical Journal, (2024) Vol. 15, Iss. 1, P. 8–22.

[18] Ibragimov G. I., Optimal pursuit time for a differential game in the Hilbert space l_2. ScienceAsia, (2013) Vol. 39S, P. 25–30.

[19] Ibragimov G. I., Ferrara M., Alias I. A., Salimi M., Ismail N., Pursuit and evasion games for an infinite system of differential equations. Bulletin of the Malaysian Mathematical Sciences Society, (2021) Vol. 45, P. 69–81.

[20] Satimov N. Y., Tukhtasinov M., On some game problems for first-order controlled evolution equations. Differential Equations, (2005) Vol. 41, Iss. 8, P. 1169–1177.

[21] Satimov N. Y., Tukhtasinov M., Game problems on a fixed interval in controlled first-order evolution equations. Mathematical Notes, (2006) Vol. 80, Iss. 3-4, P. 578–589.

[22] Ibragimov G. I., Salleh Y., Simple motion evasion differential game of many pursuers and one evader with integral constraints on control functions of players. Journal of Applied Mathematics, (2012), P. 748096.

[23] Kuchkarov A. S., Ibragimov G. I., Ferrara M., Simple motion pursuit and evasion differential games with many pursuers on manifolds with Euclidean metric. Discrete Dynamics in Nature and Society, (2016) Vol. 2016, P. 1386242.

[24] Ibragimov G. I., A problem of optimal pursuit in systems with distributed parameters. Journal of Applied Mathematics and Mechanics, (2002) Vol. 66, Iss. 5, P. 719–724.

[25] Ibragimov G., Qushaqov X., Muxammadjonov A., Pansera B. A., Guaranteed pursuit and evasion times in a differential game for an infinite system in Hilbert space l_2. Mathematics, (2023) Vol. 11, P. 2662.

[26] Ibragimov G. I., Qo'shaqov X. S., Muxammadjonov A. A., Optimal pursuit differential game problem for an infinite system of binary differential equations. Lobachevskii Journal of Mathematics, (2024) Vol. 45, Iss. 3, P. 1131–1144.

Ibragimov G.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan
University of Public Safety of the Republic of Uzbekistan,
Tashkent, Uzbekistan
e-mail: ibragimov.math@gmail.com

Muxammadjonov A.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan
e-mail: akbarshohmuhammadjonov6764@gmail.com

Qushaqov X.Sh.,
Andijan State University, Andijan, Uzbekistan.
e-mail: xolmurodjonqoshaqov737@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 90–101
DOI: 10.29229/uzmj.2026-2-11

Evasion differential game of one evader and multiple pursuers with changing energies

Ibragimov G.I., Inomiddinov S.N., Ibrokhimjonov I.I.

Abstract. In this paper, we study a simple motion evasion differential game of multiple pursuers and one evader. The control functions of the players are subject to integral constraints, where the energies of the players change linearly. If the state of the evader does not coincide with those of the pursuers forever, we say that evasion is possible in the game. We prove a theorem about evasion and construct a strategy for the evader that guarantees evasion in the game.

Keywords: differential game, integral constraint, pursuer, evader, evasion strategy, pursuit, evasion.
MSC (2020): 49N70, 49N75, 93C15

1. Introduction

During the latter half of the previous century, the theory of differential games experienced remarkable development (see, for example, [1], [2]). Within this field, multiple-pursuer differential games constitute one of the most actively investigated areas. In recent years, a substantial body of research has been devoted to this topic due to its broad range of practical applications. Differential games involving several pursuers are closely connected to real-world situations, such as reach-avoid problems, where multiple pursuers attempt to capture an evader while simultaneously reaching a target region and avoiding specific obstacles or restrictions. They are also relevant to military defense, where missile systems aim to neutralize threats while maintaining the safety of allied assets, and to surveillance and security tasks, where drones monitor a region, protecting restricted zones from intrusion.

Differential games are generally classified according to the nature of constraints imposed on players' control functions. Two fundamental categories are geometric constraint games and integral constraint games. Geometric constraints limit the control parameters to certain subsets of $\mathbb{R}^n$, often representing physical limitations such as velocity or turning rate. Most studies in the literature have concentrated on multi-player games with geometric constraints, which are briefly reviewed below.

A differential game known as the multi-pursuer, single-evader Life-line game was analyzed by Azamov [3], [4]. In this framework, the evader aims to reach a designated "life-line" (safe zone) before capture, while pursuers strive to intercept it beforehand. The author derived a winning condition expressed through Apollonius circles—if the condition is met, the pursuers succeed; otherwise, the evader prevails. Wang et al. [5] studied a differential game involving multiple pursuers and a fast evader, proposing an initial state configuration based on Apollonius circle theory to guarantee capture.

Among the most discussed topics in multi-pursuer, single-evader settings are reach-avoid planar games, examined in works such as those by Deng et al. [6], Garcia et al. [7], and Yan et al. [8]. Von Moll et al. [9] investigated a border defense scenario where multiple unmanned aerial vehicles serve as pursuers against a ground vehicle attempting to cross a boundary. Their study modeled the system as a zero-sum differential game with simple motion dynamics and derived optimal pursuit strategies from geometric principles. Similarly, Pachter [10] employed a geometric framework for single-evader capture, while Souidi et al. [11] introduced a dynamic elimination approach to enhance decision-making in multi-pursuer, multi-evader games.

In evasion games, attention shifts to the evader's strategy to avoid capture by multiple pursuers. Ibragimov [12] analyzed a game involving one fast evader attempting to escape from several pursuers and proposed a new evasion strategy ensuring escape within a region surrounding the ray emanating from the evader's initial position.

Several other contributions [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] have also explored multiple-pursuer differential games. In particular, Pshenichnii [24] examined a simple-motion differential game with multiple pursuers and a single evader, deriving pursuit and evasion

conditions based on the convex hull of the pursuers' initial states, under the assumption of identical dynamic capabilities for all players.

When such games are extended to manifolds, the dynamics of pursuit and evasion acquire richer geometric complexity. A typical example is a group of autonomous robots navigating hilly terrain to capture a target. Kuchkarov et al. [25] studied a simple-motion differential game of multiple pursuers and one evader on manifolds endowed with the Euclidean metric. They established a condition under which pursuit is guaranteed; if the condition fails, the evader succeeds in escaping.

Another important subfield with practical significance concerns multiple-capture game problems, where a group of defenders (pursuers) attempts to capture one or more attackers (evaders) m, $m \geq 1$, times. If fewer than m pursuers reach the attacker, the defenders are considered neutralized. Research on this topic includes works by Blagodatskikh and Petrov [26], Petrov [27], Shaunak and Subhash [28], Sakharov [29], and Petrov and Solov'eva [30]. When the control processes involve limited or exhaustible resources (e.g., fuel or energy), integral constraints naturally arise. A missile constrained by finite fuel reserves or a drone with limited battery capacity exemplifies this class of problems. Integral constraints regulate the total usage of a control input (such as energy or effort) throughout the game. Ibragimov and Salleh [31] investigated an evasion differential game involving multiple pursuers and a single evader under integral constraints, constructing an evasion strategy that exploits the evader's advantage in control resources. Similarly, Salimi and Ferrara [32] considered an optimal approach game with integral constraints.

In another study, Kuchkarov et al. [33] analyzed an optimal approach differential game involving several pursuers and one evader with integral constraints. The game duration was assumed to be fixed, and the payoff function was defined as the minimum distance between the evader and pursuers at the terminal time. An estimate for this payoff functional was obtained.

In the present paper, we study a simple motion evasion differential game of multiple pursuers and one evader. The control functions of the players are subject to integral constraints, where the control resources of the players change linearly. We construct an evasion strategy for the evader, which guarantees evasion in the game.

2. Statement of problem

Let the dynamics of pursuers x_i and evader y be described by the following equations:

$$\begin{array}{ll} \dot{x}_i = u_i, & x_i(0) = x_{i0}, \quad i = 1, 2, ..., M, \\ \dot{y} = v, & y(0) = y_0, \end{array} \tag{2.1}$$

where $x_i, x_{i0}, u_i, y, y_0, v \in \mathbb{R}^n$, $x_{i0} \neq y_0$, u_i and v are control parameters of the pursuers and evader, respectively.

Definition 2.1. Measurable functions $u_i(t)$ and $v(t)$, $t \geq 0$, are called admissible controls of the pursuer x_i and evader y, respectively, if they satisfy the following integral constraints:

$$\int_0^t |u_i(s)|^2 ds \leq \rho_i^2 t, \ i = 1, ..., M, \quad \int_0^t |v(s)|^2 ds \leq \sigma^2 t. \tag{2.2}$$

Definition 2.2. A function $V = (V_1, \ldots, V_n)$,

$$\big(t, y(\cdot)|_{[0,t]}, x_1(\cdot)|_{[0,t]}, \ldots, x_M(\cdot)|_{[0,t]}, u_1, \ldots, u_M\big) \mapsto V\big(t, y(\cdot)|_{[0,t]}, x_1(\cdot)|_{[0,t]}, \ldots, x_M(\cdot)|_{[0,t]}, u_1, \ldots, u_M\big),$$

is called a strategy of the evader y if for any admissible controls of the pursuers $u_i = u_i(t)$, $i = 1, \ldots, M$, the initial value problem

$$\begin{array}{ll} \dot{x}_1 = u_1(t), & x_1(0) = x_1^0, \\ \vdots & \vdots \\ \dot{x}_M = u_M(t), & x_M(0) = x_M^0, \\ \dot{y} = V\big(t, y(\cdot)|_{[0,t]}, x_1(\cdot)|_{[0,t]}, \ldots, x_M(\cdot)|_{[0,t]}, u_1, \ldots, u_M\big), & y(0) = y^0, \end{array} \tag{2.3}$$

has a unique solution $(x_1(t),\dots,x_M(t),y(t))$, $t\geq 0$, and along this solution

$$\int_0^t |V\big(s,y(\cdot)|_{[0,s]},x_1(\cdot)|_{[0,s]},\dots,x_M(\cdot)|_{[0,s]},u_1(s),\dots,u_M(s)\big)|^2\,ds\leq\sigma^2 t. \tag{2.4}$$

Here, the pursuers apply arbitrary admissible controls, while the evader adheres to a prescribed strategy. By a solution of the initial value problem (2.3), we mean an $(M+1)$-tuple $(x_1(t),\dots,x_M(t),y(t))$, $t\geq 0$, with absolutely continuous components $x_i(t)$, $i=1,\dots,M$, and $y(t)$ that satisfy the initial conditions in (2.3) and differential equations in (2.3) almost everywhere on $[0,\infty)$.

Definition 2.3. Evasion in game (2.1)-(2.2) is said to be achievable if there exists a strategy V for the evader such that, regardless of the pursuers' controls, $x_i(t)\neq y(t)$ holds for all $t\geq 0$, $i=1,2,...,M$.

We can now state the evasion problem.

Problem. Find a condition for the parameters ρ_1, ..., ρ_M, and σ, which guarantee the evasion in game (2.1)-(2.2).

This is the main problem to be investigated in this paper. It should be noted that, in the evasion game, the pursuers use arbitrary admissible controls $u_1(t),\dots,u_M(t)$, $t\geq 0$, whereas the evader uses a strategy.

3. Evasion differential game of one evader and multiple pursuers

The main result of the paper is the following statement.

Theorem 3.1. *If*

$$\sum_{i=1}^{M}\rho_i^2<\sigma^2,$$

then evasion is possible in game (2.1)-(2.2).

We study the game problem in $\mathbb{R}^2$ and the general case where $n\geq 2$ follows from the solution of the problem in $\mathbb{R}^2$.

Without loss of generality, we assume that $y_0=(0,0)$ in the $\xi_1\xi_2$-coordinate system. Let $I=\{i|\ x_{i2}(0)>0,\ \ 1\leq i\leq M\}$. There is no restriction in assuming that $I=\{1,2,...,m\}$, which means that initially the first m pursuers, i.e. the points $x_{10},x_{20},...,x_{m0}$ are in the upper half plane.

Let α and a_1 be fixed numbers that satisfy the following conditions

$$0<\alpha<\min\left\{1,\frac{\sigma-\rho}{\sqrt{2}}\right\},\quad 0<a_1<\min\left\{1,\frac{\rho^2T}{32},\ \max_{1\leq i\leq M}|y_2^0-x_{i2}^0|\right\}, \tag{3.1}$$

where $\rho=\left(\sum\limits_{i=1}^{M}\rho_i^2\right)^{1/2}$. Let

$$T_0=\frac{1}{\alpha}\max_{1\leq i\leq M}|y_2^0-x_{i2}^0|,\quad T=T_0+\frac{2a_1}{\alpha}.$$

Note that this and (3.1) imply that $\alpha a_1\leq\sigma^2T$ since

$$\alpha a_1<\alpha\max_{1\leq i\leq M}|y_2^0-x_{i2}^0|=\alpha^2T_0<\alpha^2T<\frac{(\sigma-\rho)^2}{2}T<\sigma^2T.$$

We define a sequence $a_{k+1}=\beta a_k^4$, $k=1,2,...$, where

$$\beta=\min\left\{\frac{1}{2},\frac{\alpha^2}{7200\sigma^6T^3},\frac{\alpha}{12\sigma^2T}\right\}.$$

Let $\tau_0 = 0$, and let $\tau_1 > 0$ be the first time that

$$(i) \quad |x_i(\tau_1) - y(\tau_1)| = a_1,$$
$$(ii) \quad x_{i2}(\tau_1) > y_2(\tau_1),$$

for some $i \in I$. Note that such a time τ_1 may not exist. If so we'll have $x_i(t) \neq y(t)$ for all $t \geq 0$ and $i \in I$. If there are several pursuers that satisfy conditions (i) and (ii), then without loss of generality, we can assume, by relabeling if necessary, that one of such pursuers is x_1. We call τ_1 an a_1–approach time.

If τ_1, τ_2, ..., τ_{k-1} are a_1-, a_2-, ..., a_{k-1}-approach times, respectively, then we define $\tau_k > \tau_{k-1}$ to be the a_k-approach time if the following conditions are satisfied:

$$(1) \quad |x_k(\tau_k) - y(\tau_k)| = a_k,$$
$$(2) \quad x_{k2}(\tau_k) > y_2(\tau_k),$$

for some $k \in I$. If there are several pursuers that satisfy the conditions in items (1) and (2) simultaneously, then there is no loss of generality in assuming that one of these pursuers is x_k. In this way, we define the a_k–approach times τ_k, $k = 1, 2, \ldots, m_0$, where m_0 is the number (unspecified) of all approach times. For now, we do not now m_0 is finite or $m_0 = \infty$.

Note that a_k–approach times τ_k will not necessarily be defined for all the pursuers $i \in I$. We will establish that at most one approach time will be defined for each pursuer x_i, $i \in I$, and therefore we will have $m_0 \leq m$.

Let

$$\tau_k' = \tau_k + \frac{2a_k}{\alpha}, \quad k = 1, 2, \ldots, m_0.$$

Note that we have defined τ_k and τ_k' only for $k = 1, 2, \ldots, m_0$.

Let

$$J_k = \bigcup_{j=k}^{m_0} [\tau_j, \tau_j'), \quad k = 1, 2, ..., m_0,$$

where $J_{m_0+1} \neq \emptyset$ if m_0 is finite.

We define the function

$$r : [0, T] \to \{0, 1, 2, \ldots, m_0\},$$

by the following equation

$$r(t) = \begin{cases} 0, & t \in [0, T] \setminus J_1, \\ k, & t \in [\tau_k, \tau_k') \setminus J_{k+1}, \quad k = 1, 2, \ldots, m_0, \end{cases}$$

for $m_0 > 1, \quad k = 1, 2, \ldots, (m_0 - 1)$. It is not difficult to verify that this function has the following properties:

1) $r(t) = k, \quad \tau_k \leq t < \tau_k', \quad$ if $\tau_k \leq \tau_{k+1}'$,

2) $r(t) = k, \quad \tau_k \leq t < \tau_{k+1}, \quad$ if $\tau_{k+1} < \tau_k'$.

3.1. **Strategy of the Evader.** Let $u_i(t)$, $i = 1, \ldots, M$, be arbitrary admissible controls of the pursuers. We construct the strategy of the evader as follows:

$$v(t) = V_0(t) = (0, \alpha + U(t)), \quad t \in [0, T] \setminus J_1, \tag{3.2}$$
$$v(t) = V_r(t) = (V_{r1}(t), \alpha + U(t)), \quad t \in [0, T] \cap J_1, \tag{3.3}$$
$$v(t) = (0, U(t)), \quad t > T, \tag{3.4}$$

where

$$r = r(t), \quad V_k(t) = (V_{k1}(t), \alpha + U(t)), \quad \tau_k < t < \tau_k', \quad k \in I = \{1, 2, \ldots, m\},$$

$$V_{k1}(t)=\begin{cases}\alpha+|u_{k1}(t)|, & y_1(\tau_k)\geq x_{k1}(\tau_k)\\ -(\alpha+|u_{k1}(t)|), & y_1(\tau_k)<x_{k1}(\tau_k)\end{cases},\quad U(t)=\left(\sum_{i=1}^{M}u_{i2}^2(t)\right)^{1/2}.$$

To show the admissibility of strategy (3.2)–(3.4), we consider the following functions:

$$f(t)=\begin{cases}(\alpha,\alpha), & t\in[0,T]\setminus J_1,\\ (\alpha,\alpha), & t\in J_1,\\ (0,0), & t>T,\end{cases}\qquad g(t)=\begin{cases}\left(0,\left(\sum\limits_{i=1}^{M}u_{i2}^2(t)\right)^{1/2}\right), & t\in[0,T]\setminus J_1,\\ \left(|u_{r1}(t)|,\left(\sum\limits_{i=1}^{M}u_{i2}^2(t)\right)^{1/2}\right), & t\in J_1,\\ \left(0,\left(\sum\limits_{i=1}^{M}u_{i2}^2(t)\right)^{1/2}\right), & t>T.\end{cases}$$

Then, clearly,

$$\int_0^t|f(s)|^2ds\leq 2\alpha^2t \tag{3.5}$$

and

$$|g(t)|^2=u_{r1}^2(t)+\sum_{i=1}^{M}u_{i2}^2(t)\leq\sum_{i=1}^{M}u_{i1}^2(t)+\sum_{i=1}^{M}u_{i2}^2(t)=\sum_{i=1}^{M}|u_i(t)|^2. \tag{3.6}$$

In view of (3.5) and (3.6) we have

$$\begin{aligned}\left(\int_0^t|V(s)|^2ds\right)^{1/2}&=\left(\int_0^t|f(s)+g(s)|^2ds\right)^{1/2}\leq\left(\int_0^t|f(s)|^2ds\right)^{1/2}+\left(\int_0^t|g(s)|^2ds\right)^{1/2}\\ &\leq\sqrt{2\alpha^2t}+\left(\int_0^t\sum_{i=1}^{M}|u_i(s)|^2ds\right)^{1/2}\leq\alpha\sqrt{2t}+\left(\sum_{i=1}^{M}\rho_i^2t\right)^{1/2}=\alpha\sqrt{2t}+\rho\sqrt{t}\leq\sigma\sqrt{t}\end{aligned}$$

since by choice $\alpha<\frac{\sigma-\rho}{\sqrt{2}}$. Hence, strategy (3.2)–(3.4) is admissible.

Lemma 3.2. *Let the evader apply the strategy* (3.2), (3.3), (3.4).

1) *For any* $k\in\{1,2,3,\dots,m_0\}$, *we have (i)* $\tau_k\leq T_0$, *(ii)* $\tau_k'\leq T$.
2) *If* $y_2^0\geq x_{i_02}^0$ *for some* $i_0\in\{1,2,\dots,M\}$, *then* $y_2(t)>x_{i_02}(t)$, *for all* $t>0$.

Proof. The proof of item 1(i). First, we establish $y_2(T_0)\geq x_{i2}(T_0)$ for all $i=1,2,\dots,M$. Indeed, by (3.2) and (3.3) we have, for any $i\in\{1,2,\dots,M\}$,

$$v_2(t)=\alpha+U(t)\geq\alpha+|u_{i2}(t)|,\quad 0<t\leq T_0<T, \tag{3.7}$$

since

$$U(t)=\left(\sum_{j=1}^{M}u_{j2}^2(t)\right)^{1/2}\geq|u_{i2}(t)|,\quad i\in\{1,2,\dots,M\}.$$

Hence,

$$\frac{d}{dt}(y_2(t)-x_{i2}(t))=v_2(t)-u_{i2}(t)\geq\alpha>0,\quad 0\leq t\leq T,$$

which means $y_2(t)-x_{i2}(t)$ increases strictly.

In view of $T_0 = \frac{1}{\alpha} \max\limits_{1\le i\le M} |y_2^0 - x_{i2}^0|$ and (3.7), we obtain the following:

$$\begin{aligned} y_2(T_0) - x_{i2}(T_0) &= y_2^0 + \int\limits_0^{T_0} v_2(s)ds - x_{i2}^0 - \int\limits_0^{T_0} u_{i2}(s)ds = y_2^0 - x_{i2}^0 + \int\limits_0^{T_0} (v_2(s) - u_{i2}(s))ds \\ &= y_2^0 - x_{i2}^0 + \alpha T_0 = y_2^0 - x_{i2}^0 + \max\limits_{1\le i\le M} |y_2^0 - x_{i2}^0| \ge 0. \end{aligned}$$

Consequently, $y_2(T_0) \ge x_{i2}(T_0)$, for all $i = 1, 2, \ldots, M$, which means the evader at the time T_0 is above the horizontal lines passing through the pursuers' states at the time T_0.

Next, since

$$v_2(t) = \alpha + U(t) \ge \alpha + |u_{i2}(t)|, \quad T_0 \le t \le T,$$

and

$$v_2(t) = U(t) \ge |u_{i2}(t)|, \quad t > T,$$

then we have, for all $t > T_0$, that

$$\begin{aligned} y_2(t) - x_{i2}(t) &= y_2(T_0) + \int\limits_{T_0}^{t} v_2(s)ds - x_{i2}(T_0) - \int\limits_{T_0}^{t} u_{i2}(s)ds \\ &= y_2(T_0) - x_{i2}(T_0) + \int\limits_{T_0}^{t} (v_2(s) - u_{i2}(s))ds \\ &\ge \int\limits_{[T_0,t]\cap[T_0,T]} (\alpha + |u_{i2}(s)| - u_{i2}(s))ds \\ &\ge \int\limits_{[T_0,t]\cap[T_0,T]} \alpha ds = \alpha \ \min\{t - T_0, T - T_0\} > 0. \end{aligned}$$

Hence, $y_2(t) > x_{i2}(t)$, for all $t > T_0$ and $i = 1, 2, \ldots, M$. This implies that there is no an approach time τ_k after time T_0 since by the definition of the a_k-approach time of the evader to a pursuer x_i the inequality $x_{i2}(\tau_k) > y_2(\tau_k)$ is required. However, this inequality is not satisfied after time T_0. The proof of the inequality $\tau_k \le T_0$ (item 1(i)) is complete.

Next, to prove the inequality $\tau_k' \le T$, we observe that

$$\tau_k' = \tau_k + \frac{2a_k}{\alpha} \le T_0 + \frac{2a_1}{\alpha} = T,$$

and the proof of item 1(ii) follows.

Finally, to prove item 2, we assume that $y_2^0 \ge x_{i_0 2}^0$ for some $i_0 \in \{1, 2, \ldots, M\}$. Then, clearly, for all $t > 0$,

$$\begin{aligned} y_2(t) - x_{i_0 2}(t) &= y_2^0 - x_{i_0 2}^0 + \int\limits_0^t (v_2(s) - u_{i_0 2}(s))\, ds \\ &\ge \int\limits_{[0,t]\cap[0,T]} (\alpha + |u_{i_0 2}(s)| - u_{i_0 2}(s))\, ds \ge \alpha \ min\{t, T\} > 0. \end{aligned}$$

From this we conclude that $y_2(t) > x_{i_0 2}(t)$ for all $t > 0$, which is the desired result.

From the proved lemma we draw an important conclusion that there is no approach time after T_0 meaning that all the pursuers are "behind the evader" for $t > T_0$, which implies that no pursuer can reach the evader after time T_0.

3.2. **The distance between the pursuer and evader.** Let τ_p be the a_p-approach time of a pursuer x_p, $p \in \{1, 2, \dots, m_0\}$, to the evader y.

A. To estimate the distance between $x_p(t)$ and $y(t)$, we introduce a fictitious evader (FE) whose dynamics is as follows

$$\dot{z}_p = w_p, \quad z_p(\tau_p) = y(\tau_p),$$

on the time interval $[\tau_p, \tau_p')$. Note that the initial state $z_p(\tau_p) = (z_{p1}(\tau_p), z_{p2}(\tau_p))$ of FE coincides with that $y(\tau_p) = (y_1(\tau_p), y_2(\tau_p))$ of the evader. The control w_p of FE is defined by the equation

$$w_p(t) = V_p(t) = (V_{p1}(t), \alpha + U(t)), \quad \tau_p \le t \le \tau_p'.$$

Clearly, $V_{p2}(t) = \alpha + U(t)$, and so, for the point $z_p(t) = (z_{p1}(t), z_{p2}(t))$, we have

$$z_{p2}(t) = z_{p2}(\tau_p) + \int\limits_{\tau_p}^{t} (\alpha + U(s))\, ds = y_2(\tau_p) + \int\limits_{\tau_p}^{t} v_2(s)\, ds = y_2(t), \quad \tau_p \le t \le \tau_p',$$

since $\tau_p' < T$, and so $v_2(s) = \alpha + U(s)$ by (3.2)-(3.3).

Consequently, FE $z_p(t)$, for all $\tau_p \le t \le \tau_p'$, is on the same horizontal line as the real evader $y(t)$.

For definiteness, we assume that $y_1(\tau_p) \ge x_{p1}(\tau_p)$. Then, we have

$$V_{p1}(t) = \begin{cases} \alpha + |u_{p1}(t)|, & y_1(\tau_p) \ge x_{p1}(\tau_p) \\ -(\alpha + |u_{p1}(t)|), & y_1(\tau_p) < x_{p1}(\tau_p) \end{cases} = \alpha + |u_{p1}(t)|.$$

B. We estimate the distance between $x_p(t)$ and $z_p(t)$ for $\tau_p \le t \le \tau_p'$. We estimate this distance in two ways. In the first way, using the inequality $y_1(\tau_p) \ge x_{p1}(\tau_p)$ we have the following estimate for the distance

$$\begin{aligned} |z_p(t) - x_p(t)| \ge z_{p1}(t) - x_{p1}(t) &= y_1(\tau_p) + \int\limits_{\tau_p}^{t} V_{p1}(s)\, ds - x_{p1}(\tau_p) - \int\limits_{\tau_p}^{t} u_{p1}(s) ds \\ &= y_1(\tau_p) - x_{p1}(\tau_p) + \int\limits_{\tau_p}^{t} (V_{p1}(s) - u_{p1}(s))\, ds \\ &\ge \int\limits_{\tau_p}^{t} (\alpha + |u_{p1}(s)| - u_{p1}(s))\, ds \ge \alpha(t - \tau_p) = h_1(t),\ \tau_p \le t \le \tau_p'. \end{aligned} \tag{3.8}$$

In the second way, we have

$$|z_p(t) - x_p(t)| \ge |z_p(\tau_p) - x_p(\tau_p)| - \int\limits_{\tau_p}^{t} |V_p(s) - u_p(s)|\, ds = a_p - \int\limits_{\tau_p}^{t} |V_p(s) - u_p(s)|\, ds,\ \tau_p \le t \le \tau_p'. \tag{3.9}$$

Since

$$\begin{aligned} \int\limits_{\tau_p}^{t} |V_p(s) - u_p(s)|\, ds &\le \left(\int\limits_{\tau_p}^{t} 1^2 ds \right)^{1/2} \left(\int\limits_{\tau_p}^{t} (V_p(s) - u_p(s))^2 ds \right)^{1/2} \\ &\le \sqrt{t - \tau_p} \cdot \left(\int\limits_{\tau_p}^{t} 2\left(|V_p(s)|^2 + |u_p(s)|^2 \right) ds \right)^{1/2}, \end{aligned}$$

therefore using the inequalities (recall $t \le \tau_p' < T$)

$$\int_{\tau_p}^{t} |V_p(s)|^2 ds \le \sigma^2 T, \quad \int_{\tau_p}^{t} |u_p(s)|^2 ds \le \rho^2 T < \sigma^2 T$$

we obtain

$$\int_{\tau_p}^{t} |V_p(s) - u_p(s)| ds \le \sqrt{t-\tau_p} \cdot \sqrt{2 \cdot 2\sigma^2 T} = 2\sigma\sqrt{T(t-\tau_p)}.$$

Using this and (3.9) we obtain the following

$$|z_p(t) - x_p(t)| \ge a_p - 2\sigma\sqrt{T(t-\tau_p)} = h_2(t), \ \tau_p \le t \le \tau_p'. \tag{3.10}$$

Combining (3.8) and (3.10) we obtain

$$|z_p(t) - x_p(t)| \ge h(t) = \max\{h_1(t), h_2(t)\}, \quad t \ge \tau_p. \tag{3.11}$$

We observe that the function $h_1(t) = \alpha(t-\tau_p)$, $t \ge \tau_p$, increases and the function $h_2(t) = a_p - 2\sigma\sqrt{T(t-\tau_p)}$, $t \ge \tau_p$, decreases. It is not difficult to see that the function $h(t)$, $t \ge \tau_p$, takes its minimum at the point $t = t_*$ where the graphs of the functions $h_1(t)$ and $h_2(t)$ intersect. To find this point, we solve the equation $h_1(t) = h_2(t)$:

$$\alpha(t-\tau_p) = a_p - 2\sigma\sqrt{T(t-\tau_p)}.$$

Denote $\sqrt{t-\tau_p} = d$ to obtain the equation

$$\alpha d^2 + 2\sigma\sqrt{T}d - a_p = 0. \tag{3.12}$$

Using the inequality $\alpha a_p < \alpha a_1 \le \sigma^2 T$, we estimate the positive root of equation (3.12):

$$d = \frac{-\sigma\sqrt{T} + \sqrt{\sigma^2 T + \alpha a_p}}{\alpha} = \frac{a_p}{\sigma\sqrt{T} + \sqrt{\sigma^2 T + \alpha a_p}} \ge \frac{a_p}{\sigma\sqrt{T} + \sqrt{\sigma^2 T + \sigma^2 T}} = \frac{a_p}{\sigma\sqrt{T}(1+\sqrt{2})}.$$

Hence,

$$h(t_*) = h_1(t_*) = \alpha d^2 = \frac{\alpha a_p^2}{\left(\sigma\sqrt{T}(1+\sqrt{2})\right)^2} \ge \frac{\alpha a_p^2}{6\sigma^2 T}. \tag{3.13}$$

Since the FE moves only on the time interval $[\tau_p, \tau_p']$, therefore, we obtain from (3.11) and (3.13) that

$$|z_p(t) - x_p(t)| \ge h(t_*) > \frac{\alpha a_p^2}{6\sigma^2 T}, \quad \tau_p \le t \le \tau_p'. \tag{3.14}$$

C. We estimate the distance between the FE $z_p(t)$ and the evader $y(t)$ on $\tau_p \le t \le \tau_p'$.

Using equation $z_p(\tau_p) = y(\tau_p)$ we have the following

$$|y(t) - z_p(t)| = \left| \int_{\tau_p}^{t} (v(s) - V_p(s))\, ds \right|, \quad \tau_p \le t \le \tau_p', \tag{3.15}$$

where $V_p(t) = (V_{p1}(t), \alpha + U(t))$, $\tau_p \le t < \tau_p'$.

To estimate (3.15), we consider the following two cases: (i) $\tau_{p+1} \ge \tau_p'$ (ii) $\tau_{p+1} < \tau_p'$.

If (i) $\tau_{p+1} \ge \tau_p'$ is the case, then by the property of the function $r(t)$ we have $r = r(t) = p$ for $\tau_p \le t < \tau_p'$, and hence, by (3.3) the maneurer $v(t) = V_r(t) = (V_{r1}(t), \alpha + U(t))$ has the form $v(t) = V_p(t)$, $\tau_p \le t < \tau_p'$. Then, clearly, (3.15) implies that

$$|y(t) - z_p(t)| = 0.$$

Let now (ii) $\tau_{p+1} \le \tau'_p$ is the case. Then, $r = r(t) = p$ for $\tau_p \le t < \tau_{p+1}$, and, hence,

$$v(t) = V_p(t), \quad \tau_p \le t < \tau_{p+1}.$$

Consequently, $v(t)$ coincides with $w_p(t)$ on $\tau_p \le t < \tau_{p+1}$, implying that $y(t) = z_p(t)$, $t \in [\tau_p, \tau_{p+1}]$. In particular, we have $y(\tau_{p+1}) = z_p(\tau_{p+1})$. Using this we obtain the following, for the distance $|y(t)-z_p(t)|$ on $[\tau_{p+1}, \tau'_p]$,

$$|y(t) - z_p(t)| = \left| \int\limits_{\tau_{p+1}}^{t} (v(s) - V_p(s))ds \right| \le \int\limits_{\tau_{p+1}}^{t} |v(s) - V_p(s)|ds$$
$$\le \int\limits_{[\tau_{p+1},t]\setminus J_{p+1}} |v(s) - V_p(s)|ds + \int\limits_{[\tau_{p+1},t]\cap J_{p+1}} |v(s) - V_p(s)|ds. \tag{3.16}$$

Since $r(t) = p$ for $t \in [\tau_p, \tau'_p) \setminus J_{p+1}$ and $[\tau_{p+1}, t] \setminus J_{p+1} \subseteq [\tau_p, \tau'_p) \setminus J_{p+1}$, therefore $v(t) = V_p(t)$ for $t \in [\tau_{p+1}, t] \setminus J_{p+1}$. This implies that the first integral in (3.16) is equal to 0. Thus, (3.16) takes the form

$$|y(t) - z_p(t)| \le \int\limits_{[\tau_{p+1},t]\cap J_{p+1}} |v(s) - V_p(s)|\, ds.$$

From this in view of

$$|v(s) - V_p(s)| = |V_{r1}(s) - V_{p1}(s)| \le |V_{r1}(s)| + |V_{p1}(s)| = 2\alpha + |u_{r1}(s)| + |u_{p1}(s)|$$

we obtain

$$|y(t) - z_p(t)| \le \int\limits_{J_{p+1}} (2\alpha + |u_{r1}(s)| + |u_{p1}(s)|)\, ds \quad = \int\limits_{J_{p+1}} 2\alpha\, ds + \int\limits_{J_{p+1}} |u_{r1}(s)|\, ds + \int\limits_{J_{p+1}} |u_{p1}(s)|\, ds. \tag{3.17}$$

To estimate the first integral in (3.17), we have the following

$$\int\limits_{J_{p+1}} 2\alpha\, ds \le 2\alpha\, \mathrm{mes}(J_{p+1}) \le 2\alpha \sum_{j=p+1}^{m_0} (\tau'_j - \tau_j) \le 2\alpha \sum_{j=p+1}^{m_0} \frac{2a_j}{\alpha} \le 4 \sum_{j=p+1}^{\infty} a_j \le 8a_{p+1}. \tag{3.18}$$

To estimate the second and third integrals in (3.17), we use the Cauchy-Schwarz inequality and (3.18) to obtain

$$\int\limits_{J_{p+1}} |u_{r1}(s)|ds \le \left(\int\limits_{J_{p+1}} 1^2 ds \right)^{1/2} \left(\int\limits_{Jp+1} |u_{r1}(s)|^2(s)ds \right)^{1/2} \le \sqrt{8a_{p+1}}\ \rho\sqrt{T} \tag{3.19}$$

and

$$\int\limits_{J_{p+1}} |u_{p1}(s)|ds \le \left(\int\limits_{J_{p+1}} 1^2 ds \right)^{1/2} \left(\int\limits_{Jp+1} |u_{p1}(s)|^2(s)ds \right)^{1/2} \le \sqrt{8a_{p+1}}\ \rho\sqrt{T}. \tag{3.20}$$

Combining equations (3.17), (3.18), (3.19) and (3.20) we obtain the following estimate for $|y(t)-z_p(t)|$:

$$|y(t) - z_p(t)| \le 8a_{p+1} + 4\rho\sqrt{T}\sqrt{2a_{p+1}} \le 5\rho\sqrt{2T}\sqrt{a_{p+1}}. \tag{3.21}$$

Here we used the inequality $a_{p+1} < a_1 < \frac{\rho^2 T}{32}$ following from (3.1).

D. We estimate the distance between the pursuer $x_p(t)$ and the evader $y(t)$ on the time interval $[\tau_p, \tau_p']$. By (3.14) and (3.21) $a_{p+1} \le \frac{\alpha^2}{7200\sigma^6T^3}a_p^4$ we have

$$\begin{aligned}|y(t)-x_p(t)| &\ge |x_p(t)-z_p(t)| - |z_p(t)-y(t)|\\ &\ge \frac{\alpha a_p^2}{6\sigma^2 T} - 5\rho\sqrt{2T}\sqrt{a_{p+1}} \ge \frac{\alpha a_p^2}{12\sigma^2 T}.\end{aligned} \tag{3.22}$$

Next, we estimate $y_2(t) - x_{p2}(t)$ for $t > \tau_p'$. We observe

$$y_2(\tau_p) - x_{p2}(\tau_p) \ge -|y(\tau_p) - z_p(\tau_p)| = -a_p.$$

Since $z_{p2}(\tau_p) = y_2(\tau_p)$, we have

$$\begin{aligned}z_{p2}(\tau_p') - x_{p2}(\tau_p') &= y_2(\tau_p) - x_{p2}(\tau_p) + \int_{\tau_p}^{\tau_p'} (\alpha + U(s) - u_{p2}(s))\, ds\\ &\ge -a_p + \int_{\tau_p}^{\tau_p'} \left(\alpha + \left(\sum_{i=1}^{M} u_{i2}^2(s)\right)^{1/2} - u_{p2}(s)\right) ds\\ &\ge -a_p + \alpha \int_{\tau_p}^{\tau_p'} ds = -a_p + \alpha(\tau_p' - \tau_p) = a_p.\end{aligned} \tag{3.23}$$

Here we used the definition $\tau_p' = \tau_p + \frac{2a_p}{\alpha}$.

By definition of control of FE and evader $v_2(t) = V_{p2}(t) = \alpha + U(t)$, $t \in [\tau_p, \tau_p')$ and therefore $y_2(t) = z_{p2}(t)$, $t \in [\tau_p, \tau_p']$. In particular, $y_2(\tau_p') = z_{p2}(\tau_p')$. In addition, since $v_2(t) = \alpha + U(t)$, $t \in [0, T]$ and $v_2(t) = U(t)$, $t > T$, therefore, we have $v_2(t) \ge U(t) \ge |u_{p2}(t)|$ for all $t \ge 0$. Hence, by (3.23), for $t > \tau_p'$,

$$\begin{aligned}y_2(t) - x_{p2}(t) &= y_2(\tau_p') + \int_{\tau_p'}^{t} v_2(s)ds - \left(x_{p2}(\tau_p') + \int_{\tau_p'}^{t} u_{p2}(s)ds\right)\\ &= z_{p2}(\tau_p') - x_{p2}(\tau_p') + \int_{\tau_p'}^{t} (v_2(s) - u_{p2}(s))\, ds \ge z_{p2}(\tau_p') - x_{p2}(\tau_p') \ge a_p.\end{aligned} \tag{3.24}$$

Consequently,

$$y_2(t) - x_{p2}(t) \ge a_p, \quad t > \tau_p'. \tag{3.25}$$

In summary,

1) $|y(t) - x_p(t)| > a_p$, $0 \le t < \tau_p$, by the definition of time τ_p,
2) $|y(t) - x_p(t)| \ge \frac{\alpha a_p^2}{12\sigma^2 T} > a_{p+1}$, $\tau_p \le t \le \tau_p'$, by (3.22) and inequalities
$$a_{p+1} \le \frac{\alpha}{12\sigma^2 T}a_p^4 < \frac{\alpha}{12\sigma^2 T}a_p^2,$$
3) $|y(t) - x_p(t)| \ge a_p$, $t > \tau_p'$, by (3.25).

These inequalities allow us to draw the conclusion that $x_p(t) \ne y(t)$ for all $t \ge 0$, which means that evasion is possible.

The conclusions 1)-3) derived above imply that $|y(t) - x_p(t)| \ge a_{p+1}$ for the pursuer x_p, for which the a_p-approach time τ_p is occurred. This means that for the pursuer x_p another a_k-approach time with $k \ge p+1$ further will not occur. Therefore, for each pursuer x_i, $i \in \{1, 2, ..., m\}$, either an a_i-approach time τ_i will not occur or it occurs only once. Hence, the number m_0 of all approach times τ_i is finite. More precisely, $m_0 \le m$. The proof of the theorem is complete.

4. Conclusions

We have studied a differential game of evasion with multiple pursuers and a single evader in which all players' control resources vary over time. The evader's available control resource exceeds the aggregate resource of all pursuers. To the best of our knowledge, evasion problems under integral constraints of this type have not been addressed in the literature.

Our main contributions are (i) an evasion theorem establishing sufficient conditions for successful evasion and (ii) an explicit evasion strategy for the evader that guarantees capture avoidance under those conditions.

As future work, it would be natural to investigate analogous games under alternative constraint classes for the evader and/or to extend the analysis to higher-order dynamics.

5. Acknowledgment

This research was carried out under Research Project No. FL-9524115148 of the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan.

References

[1] Chikrii A. A., Conflict-Controlled Processes. Kluwer Academic Publishers Group, Dordrecht, (1997).

[2] Petrosyan L. A., Differential Games of Pursuit. World Scientific, Singapore, (1993).

[3] Azamov A. A., On the quality problem for simple pursuit games with constraint. Serdica Bulg. Math. Publ. Sofia, (1986) Vol. 12, P. 38–43.

[4] Azamov A., Ibragimov G., Kurbanov A., Evasion differential game of one faster evader with bounded maneuverability from multiple pursuers. Dynamic Games and Applications, (2025),

[5] Wang Q., Fei Q., Wang B., Cooperative hunting strategy with multi-pursuer and a superior evader. China Automation Congress (CAC), Chongqing, China, (2023), P. 6388–6393.

[6] Deng R., Zhang W., Yan R., Shi Z., Zhong Y., Multiple-pursuer single-evader reach avoid games in constant flow fields. IEEE Transactions on Automatic Control, (2024) Vol. 69, P. 1789–1795.

[7] Garcia E., Casbeer D. W., Pachter M., Optimal strategies for a class of multi-player reach-avoid differential games in 3D space. IEEE Robotics and Automation Letters, (2020) Vol. 5, P. 4257–4264.

[8] Yan R., Deng R., Duan X., Shi Z., Zhong Y., Multiplayer reach-avoid differential games with simple motions: A review. Frontiers in Control Engineering, (2023) Vol. 3, P. 1093186.

[9] Von Moll A., Garcia E., Casbeer D., Suresh M., Swar S. C., Multiple-pursuer, single-evader border defense differential game. Journal of Aerospace Information Systems, (2020) Vol. 17, P. 407.

[10] Pachter M., Von Moll A., Garcia E., Casbeer D., Milutinovic D., Cooperative pursuit by multiple pursuers of a single evader. Journal of Aerospace Information Systems, (2020) Vol. 17, P. 371–389.

[11] Souidi M. E. H., Maarouk T. M., Ledmi M., Ledmi A., Rahab H., Multi-pursuer multi evader games based on dynamic elimination priorities. Journal of Computer and Systems Sciences International, (2023) Vol. 62, P. 398–411.

[12] Ibragimov G. I., Evasion differential game of one evader and many slow pursuers. Dynamic Games and Applications, (2024) Vol. 14, P. 665–685.

[13] Fang X., Wang C., Xie L., Chen J., Cooperative pursuit with multi-pursuer and one faster free-moving evader. IEEE Transactions on Cybernetics, (2022) Vol. 52, P. 1405–1414.

[14] Selvakumar J., Bakolas E., Feedback strategies for a reach-avoid game with a single evader and multiple pursuers. IEEE Transactions on Cybernetics, (2021) Vol. 51, P. 696–707.

[15] Makkapati V. R., Tsiotras P., Optimal evading strategies and task allocation in multiplayer pursuit-evasion problems. Dynamic Games and Applications, (2019) Vol. 9, P. 1168–1187.

[16] Scott W. L., Leonard N. E., Optimal evasive strategies for multiple interacting agents with motion constraints. Automatica, (2018) Vol. 94, P. 26–34.

[17] Lopez V. G., Lewis F. L., Wan Y., Sanchez E. N., Fan L., Solutions for multiagent pursuit evasion games on communication graphs. IEEE Transactions on Automatic Control, (2020) Vol. 65, P. 1911–1923.

[18] Cheng L., Yuan Y., Multiplayer obstacle avoidance pursuit-evasion games with adaptive parameter estimation. IEEE Transactions on Industrial Electronics, (2023) Vol. 70, P. 5171–5181.

[19] Zhou P., Chen B. M., Distributed optimal solutions for multiagent pursuit-evasion games for capture and formation control. IEEE Transactions on Industrial Electronics, (2023) Vol. 71, P. 5224–5234.

[20] Ramana M. V., Kothari M., Pursuit strategy to capture high-speed evaders using multiple pursuers. Journal of Guidance, Control, and Dynamics, (2017) Vol. 40, P. 139–149.

[21] Machtakova A. I., Petrov N. N., On two problems of pursuit of a group of evaders in differential games with fractional derivatives. Vestn. Udmurt. Univ. Mat. Mekh. Komp'yut. Nauki, (2024) Vol. 34, P. 65–79.

[22] Petrov N. N., Solov'eva N. A., Problem of group pursuit in linear recurrent differential games. Journal of Mathematical Sciences, (2018) Vol. 230, P. 732–736.

[23] Samatov B. T., Problems of group pursuit with integral constraints on controls of the players, I. Cybernetics and Systems Analysis, (2013) Vol. 49, P. 756–767.

[24] Pshenichnii B. N., Simple pursuit by several objects. Cybernetics and System Analysis, (1976) Vol. 12, P. 484–485.

[25] Kuchkarov A. S., Ibragimov G. I., Ferrara M., Simple motion pursuit and evasion differential games with many pursuers on manifolds with Euclidean metric. Discrete Dynamics in Nature and Society, (2016) Vol. 2016, P. 1386242.

[26] Blagodatskikh A. I., Petrov N. N., Simultaneous multiple capture of rigidly coordinated evaders. Dynamic Games and Applications, (2019) Vol. 9, P. 594–613.

[27] Petrov N. N., Multiple capture in a group pursuit problem with fractional derivatives and phase restrictions. Mathematics, (2021) Vol. 9, P. 1171.

[28] Shaunak D. B., Subhash S., k-Capture in multiagent pursuit evasion, or the lion and the hyenas. Theoretical Computer Science, (2014) Vol. 522, P. 13–23.

[29] Sakharov D. V., Multiple capture in Pontryagin's almost periodic example. Journal of Computer and Systems Sciences International, (2017) Vol. 56, P. 576–583.

[30] Petrov N. N., Solov'eva N. A., Multiple capture of given number of evaders in linear recurrent differential games. Journal of Optimization Theory and Applications, (2019) Vol. 182, P. 417–429.

[31] Ibragimov G. I., Salleh Y., Simple motion evasion differential game of many pursuers and one evader with integral constraints. Journal of Applied Mathematics, (2012) Vol. 2012, P. 748096.

[32] Salimi M., Ferrara M., Differential game of optimal pursuit of one evader by many pursuers. International Journal of Game Theory, (2019) Vol. 48, P. 481–490.

[33] Kuchkarov A. S., Ibragimov G. I., Khakestari M., On a linear differential game of optimal approach of many pursuers with one evader. Journal of Dynamical and Control Systems, (2013) Vol. 19, P. 1–15.

Ibragimov G.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan
University of Public Safety of the Republic of Uzbekistan,
Tashkent, Uzbekistan
e-mail: ibragimov.math@gmail.com

Inomiddinov S.N.,
Namangan State University, Namangan, Uzbekistan.
e-mail: sn.inomiddinov@gmail.com

Ibrokhimjonov I.I.,
National University of Uzbekistan, Toshkent, Uzbekistan
e-mail: islombekibrokhimjonov03@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 102–110
DOI: 10.29229/uzmj.2026-2-12

Geometric description of the unit ball of a neutral strongly facially symmetric space

Ibragimov M.

Abstract. This paper is devoted to the study of the geometric structure of neutral strongly facially symmetric (SFS) spaces, which provide a natural framework for modeling aspects of quantum mechanics in geometric terms. We investigate the representation of symmetric faces associated with complete and maximal geometric tripotents and establish their connections with split faces. Based on these results, we obtain a geometric description of the unit ball in neutral SFS-spaces. Furthermore, we give a characterization of real neutral SFS-spaces, highlighting their structural properties and the role of geometric tripotents in determining the organization of the unit ball.

Keywords: strongly facially symmetric spaces, neutral spaces, geometric tripotents, symmetric faces, split faces, unit ball, convex geometry
MSC (2020): 4B20

1. Introduction

Mathematical models of quantum mechanics provide a powerful and versatile framework for describing the behavior of quantum systems. These models comprise a wide spectrum of methods and approaches, reflecting both algebraic and geometric perspectives.

In the works of Sh.A. Ayupov, B. Iochum, and N. Yadgorov, a so-called "convex" model was developed, whose central concept is the study of convex sets of states of physical systems. Within this framework, the authors investigated geometric characterizations of projective convex sets [1, 2, 3, 4]. Further contributions by E. Alfsen, F. Schultz, H. Hanche-Olsen, and B. Iochum [5, 6] established geometric conditions on convex sets that ensure their affine isomorphism to the state space of a C^*-algebra, a von Neumann algebra, or a JB-algebra. The geometric properties of convex sets presented in the works of E. Alfsen and F. Schultz demonstrate the existence of unordered analogues of JB^*-triples.

A significant geometric approach to quantum mechanics is provided by the theory of strongly facially symmetric (SFS) spaces, introduced by J. Friedman and B. Russo in [7, 8]. These spaces supply a natural framework for analyzing the unit ball of the predual of a JBW^*-triple, enabling one to describe the associated convex sets entirely in geometric terms.

In [8, Proposition 4.5] it was proved that for any fixed geometric tripotent ω in a neutral strongly facially symmetric space Z, the set

$$L_\omega := \{v \in G\mathfrak{U} : v \leq \omega\} \cup \{0\}$$

is a complete orthomodular lattice with the least element 0, the greatest element ω, and orthocomplementation given by $v \mapsto v^\perp = \omega - v$, where $G\mathfrak{U}$ denotes the set of all geometric tripotents of the unit ball of the dual space Z^*. It should be emphasized that a detailed description and representation of the symmetric face F_ω corresponding to the geometric tripotent ω makes it possible to obtain a geometric characterization of the unit ball of a facially symmetric space and to describe its internal structure, since symmetric faces of the unit ball play a fundamental role in the theory of facially symmetric spaces.

In [9] the notion of a strongly split face of the unit ball of a neutral strongly facially symmetric space was introduced, and it was shown that if for every element $u \in L_\omega$ the symmetric face F_u is a strongly split face, then the lattice L_ω acquires the structure of a Boolean algebra.

In the present paper, relying on the indicated structural role of symmetric and strongly split faces, we describe a representation of the symmetric face F_ω corresponding to a complete (see Theorem 3.1) and maximal (see Theorem 3.4) geometric tripotent ω, by means of a split face F_u, where $u \in L_\omega$.

Such a representation makes it possible to explicitly relate the lattice structure of geometric tripotents belonging to L_ω to the geometry of the corresponding symmetric face of the unit ball of a facially symmetric space. It should be stressed that in the theory of facially symmetric spaces the unit ball is the principal geometric object, since its facial structure completely determines the geometry and analytic properties of the space. In this context, Theorems 3.1 and 3.4 play a key role, as they provide a precise description of the structure of symmetric faces corresponding to complete and maximal geometric tripotents and thereby form the basis for a geometric characterization of the unit ball. As direct consequences of these results, Corollary3.8 gives a representation of the unit ball of a neutral strongly facially symmetric space, and Corollary 3.9 provides a description of a real neutral SFS space.

Throughout this work, we adhere to the terminology and notation established in [7, 8, 10, 11, 12, 13].

2. Preliminaries

We start by reviewing certain notions from the theory of JBW^*-triples that are essential for formulating and understanding the corresponding concepts in facially symmetric spaces.

Definition 2.1. A Banach space E over $\mathbb{C}$ is called a *JB^*-triple* if it is equipped with a triple product $(a,b,c) \mapsto \{a,b,c\}$ mapping $E \times E \times E$ into E, such that

(i) $\{a,b,c\}$ is linear in a and c, and conjugate linear in b;

(ii) $\{a,b,c\}$ is symmetric in the outer variables, i.e. $\{a,b,c\} = \{c,b,a\}$;

(iii) for every $x \in E$, the operator $L(x,x) : E \to E$ defined by $L(x,x)y = \{x,x,y\}$, where $y \in E$, is Hermitian with nonnegative spectrum;

(iv) the triple product satisfies the following identity, called the *main identity*:

$$L(x,x)\{a,b,c\} = \{L(x,x)a,b,c\} - \{a,L(x,x)b,c\} + \{a,b,L(x,x)c\};$$

(v) for every $x \in E$, the following equality holds:

$$\|\{x,x,x\}\| = \|x\|^3.$$

A nonzero element u of a JB^*-triple E is called a *tripotent* if $u = \{u,u,u\}$. Two elements a and b are said to be *orthogonal* ($a \perp b$) if $\{a,b,x\} = 0$ for every $x \in E$. The *quadratic operator* Q on E is defined (see [13, §1]) by

$$Q(x)y = \{x,y,x\}, \quad x,y \in E.$$

The *Peirce projections* $P_k(u)$ $(k = 0,1,2)$ corresponding to a tripotent u are defined as

$$P_2(u) = Q(u)^2, \quad P_1(u) = 2\big(L(u,u) - Q(u)^2\big), \quad P_0(u) = I - 2L(u,u) + Q(u)^2.$$

It is clear that $\sum_{j=0}^{2} P_j(u) = I$ and $L(u,u) = P_2(u) + \frac{1}{2}P_1(u)$. Moreover, in [13, Corollary 1.2] it was shown that the operators $P_k(u)$ $(k = 0,1,2)$ are contractive projections.

We denote by $E_k(u)$ the range of the projection $P_k(u)$: $E_k(u) = P_k(u)E$, where $k = 0,1,2$. It is known (see [14, Fact 4.2.14]) that

$$E = E_2(u) \oplus E_1(u) \oplus E_0(u), \quad E_j(u) = \{x \in E : \{u,u,x\} = \tfrac{j}{2}x\}, \quad j = 0,1,2.$$

A JB^*-triple E whose predual E_* is a Banach space is called a *JBW^*-triple*.

We now introduce the key concepts from the theory of facially symmetric spaces, illustrate them with examples, and explain the motivation behind their definitions. For clarity, we compare these notions with their analogues in operator algebra theory, particularly in the theory of JBW^*-triples. This approach makes it possible to trace the logical steps leading to the definition of facially symmetric spaces and to describe the structure of their unit balls.

Let Z be a real or complex normed space. A central notion in the construction of facially symmetric spaces is the definition of orthogonality of elements of Z: elements $f, g \in Z$ are called *orthogonal* ($f \diamondsuit g$) if $\|f \pm g\| = \|f\| + \|g\|$.

Example 2.2. [15, Lemma 2]. Let f and g be normal functionals on a von Neumann algebra A, and let u and v be partial isometries corresponding to their polar decompositions, respectively. Then u and v are orthogonal if and only if $f \diamondsuit g$.

Example 2.3. Let E_* be the predual of a JBW^*-triple E. Then for every nonzero element $f \in E_*$ there exists a nonzero tripotent $u \in E$ which is the least tripotent such that

$$u(f) = \|f\|.$$

This u is called the *support tripotent* for f and is denoted by $v(f)$ (see [16, 13]).

In [17, Theorem 5.4] it was proved that for $f, g \in E_*$,

$$f \diamondsuit g \quad \Leftrightarrow \quad v(f) \perp v(g).$$

Recall that a convex subset F of the unit ball $Z_1 = \{f \in Z : \|f\| \leq 1\}$ is called a *face* if for $f, g \in Z_1$ and $t \in (0,1)$ the condition $tf + (1-t)g \in F$ implies $f, g \in F$. A face F is called a *norm-exposed face* if $F = F_u = \{f \in Z_1 : f(u) = 1\}$ for some $u \in Z^*$ with $\|u\| = 1$.

Subsets $S, T \subset Z$ are called *orthogonal* ($S \diamondsuit T$) if $f \diamondsuit g$ for all $(f, g) \in S \times T$. For a subset S of Z we set

$$S^{\diamondsuit} = \{f \in Z : \ f \diamondsuit g, \ \forall g \in S\}$$

and call $S^{\diamondsuit}$ the *orthogonal complement* of S.

An element $u \in Z^*$ is called a *projective unit* if $\|u\| = 1$ and $\langle g, u \rangle = 0$ for all $g \in F_u^{\diamondsuit}$. Note that if $F_u \neq \emptyset$, then F_u is a norm-exposed face of the unit ball Z_1 parallel to $F_u^{\diamondsuit}$, i.e. $\langle u, F_u \rangle = 1$ and $\langle u, F_u^{\diamondsuit} \rangle = 0$.

We denote by $\mathfrak{F}$ and $\mathfrak{U}$ the sets of all norm-exposed faces of Z_1 and all projective units in Z_1^*, respectively.

Definition 2.4. A norm-exposed face F is called a *symmetric face* if there exists a linear isometry $S_F : Z \to Z$ with $S_F^2 = I$ such that the fixed-point set of S_F is $\overline{sp} F \oplus F^{\diamondsuit}$. In particular, $F^{\diamondsuit}$ is a closed linear subspace. We call S_F the *symmetry* associated with F.

If a symmetric face F of the unit ball Z_1 consists of a single point, it is called a *norm-exposed point*. We denote by $\exp Z_1$ the set of all norm-exposed points of Z_1.

Using symmetric faces, we now define the notion of a geometric tripotent, which plays the role of the analogue of a tripotent in a JBW^*-triple.

Definition 2.5. A projective unit $u \in Z^*$ is called a *geometric tripotent* if F_u is a symmetric face and $S_{F_u}^* u = u$ for the symmetry S_{F_u}.

We denote by $S\mathfrak{F}$ and $G\mathfrak{U}$ the sets of all symmetric faces of Z_1 and all geometric tripotents in Z_1^*, respectively.

Definition 2.6. A real or complex normed space Z is called a *weakly facially symmetric space* (WFS-*space*) if every norm-exposed face of Z_1 is a symmetric face.

One of the fundamental examples of a WFS-space is the predual of a von Neumann algebra (see [12, Lemma 2.8 and Lemma 2.10]). It should also be noted that in [8, Proposition 1.6] the bijectivity of the mapping

$$G\mathfrak{U} \ni u \ \longmapsto \ F_u \in S\mathfrak{F}$$

was established in the case where Z is a WFS-space.

Earlier, we introduced the definitions of Peirce projections in a JBW^*-triple. A significant property of JBW^*-triples is the existence of a Peirce decomposition associated with a given norm-exposed face, where each component is itself a JBW^*-triple.

On a WFS-space the projections $P_k(F_u) \quad (k \in \{0, 1, 2\})$, which are analogues of the Peirce projections in a JBW^*-triple, are defined for each symmetric face F_u ($u \in G\mathfrak{U}$). They are called *geometric* (or *generalized*) Peirce projections:

$$P_1(F_u) = \tfrac{1}{2}(I - S_{F_u}), \quad P_1(F_u)(Z) = \{f \in Z : \ S_{F_u} f = -f\}.$$

Here $P_0(F_u)$ and $P_2(F_u)$ project Z onto $F_u^{\diamondsuit}$ and $\overline{sp}F_u$, respectively.

For convenience, we introduce the notation:

$$Z^* = U,\ \ P_k(F_u) = P_k(u),\ \ P_k(u)Z = Z_k(u),\ \ Z_k^*(u) = U_k(u),\ \ S_{F_u} = S_u,$$
$$T = \{\lambda \in K:\ |\lambda| = 1\},\quad k \in \{1,2,3\},\ \ K = \mathbb{R} \text{ or } K = \mathbb{C}.$$

It is clear that

$$P_0(u) + P_2(u) = \tfrac{1}{2}(I + S_u),\quad P_0(u) + P_1(u) + P_2(u) = I,\quad P_0(u) - P_1(u) + P_2(u) = S_u.$$

Recall that a contractive projection Q on a normed space Z is said to be *neutral* if, for every $f \in Z$, the equality $\|Qf\| = \|f\|$ implies $Qf = f$. A normed space Z is called *neutral* if, for each symmetric face F_u, the corresponding projection $P_2(u)$ is neutral.

Definition 2.7. A WFS-space Z is called a *strongly facially symmetric space* (SFS-*space*) if for each symmetric face F_u of Z_1 and for every $v \in Z^*$ with $\|v\| = 1$ and $F_u \subset F_v$, we have $S_u^* v = v$.

Important examples of SFS-spaces are the predual spaces of von Neumann algebras [12, Theorem 2.11] and the predual spaces of JBW^*-triples [12, Theorem 3.1]. In these cases, the SFS-spaces are neutral.

Definition 2.8. Let Z be a strongly facially symmetric space. A geometric tripotent u is called

- *maximal* if $U_0(u) = \{0\}$;
- *complete* if $U_2(u) = U$;
- *unitary* if $Z_1 = \mathrm{co}(F_u \cup -F_u)$.

For geometric tripotents u, v we write $u \le v$ if $F_u \subset F_v$. If $u \le v$ and $u \ne v$, we write $u < v$.

Definition 2.9. (see [7, §2]) Let Z be a normed space. Elements $u, v \in Z^*$ are called *orthogonal* if there exists a symmetric face $F \subset Z_1$ such that

(i) $u \in \operatorname{im} P_2(F)^*$ and $v \in \operatorname{im} P_0(F)^*$;

or

(ii) $u \in \operatorname{im} P_0(F)^*$ and $v \in \operatorname{im} P_2(F)^*$.

In this case we write $u \diamondsuit v$ and $v \diamondsuit u$.

Definition 2.10. A neutral SFS-space Z is said to satisfy the FE-*condition* if every norm-closed face of Z_1 (different from Z_1 itself) is a norm-exposed face.

3. Geometric description of the unit ball of a neutral strongly facially symmetric space

Recall that a face F of a convex set K is called a *split face* if there exists a face G, called the *complementary face* to F, such that $F \cap G = \emptyset$ and K is the direct convex sum $F \oplus_c G$, i.e., every element $f \in K$ can be uniquely represented in the form

$$f = tg + (1-t)h,\quad t \in [0,1],\ g \in F,\ h \in G.$$

Let Z be a strongly facially symmetric space and let $u, \omega \in G\mathfrak{U}$. If $u < \omega$, then by [8, Lemma 4.2] it follows that $(\omega - u) \in G\mathfrak{U}$ and $(\omega - u) \diamondsuit u$. In this case the geometric tripotent $\omega - u$ is called the *orthogonal complement* of u and is denoted by $u^{\perp}$.

Let Z be a neutral SFS-space and let ω be a fixed geometric tripotent in the space Z^*. A symmetric face $F_u \subset F_\omega$ is called a *strongly split face* if the following equality holds:

$$F_\omega = F_u \oplus_c F_{u^\perp}.$$

Theorem 3.1. *Let Z be a neutral SFS-space with the (FE)-condition and let ω be a complete geometric tripotent. If a symmetric face $F_u \subset F_\omega$ corresponding to a geometric tripotent u is a split face, then F_ω admits the following representation:*

$$F_\omega = \left\{ f \in Z_1:\ f = \|P_2(u)f\| g + \|P_2(u^\perp)f\| h,\ g \in F_u,\ h \in F_{u^\perp} \right\}. \tag{3.1}$$

We first prove the following auxiliary lemmas.

Lemma 3.2. *Let Z be a neutral SFS-space. If ω is a complete geometric tripotent, then for every $f \in F_\omega$ and for every $u < \omega$ we have:*

(i) $\|P_2(u)f\| + \|P_2(u^\perp)f\| = 1$;

(ii) $\|P_2(u)f\|g + \|P_2(u^\perp)f\|h \in F_{u+u^\perp}, \quad g \in F_u,\, h \in F_{u^\perp}$.

Proof: (i) From [18, Corollary 3.4(iii)] it follows that $P_2(u^\perp) = P_0(u)$. Hence, by [8, Proposition 1.5] we obtain:

$$\|P_2(u)f\| + \|P_2(u^\perp)f\| = \|P_2(u)f\| + \|P_0(u)f\| = \|P_2(u)f + P_0(u)f\| = \frac{\|f + S_u f\|}{2}.$$

According to [19, Lemma 1] we have $S_u f \in F_\omega$, hence $(f + S_u f)(\omega) = 2$, and therefore

$$\|P_2(u)f\| + \|P_2(u^\perp)f\| = \frac{\|f + S_u f\|}{2} = 1.$$

(ii) Since $g \diamondsuit h$, by [20, p. 5.127] we have

$$\|P_2(u)f\|g \diamondsuit \|P_2(u^\perp)f\|h.$$

Then, by the definition of orthogonality and property (i), it follows that

$$\big\|\|P_2(u)f\|g + \|P_2(u^\perp)f\|h\big\| = \|P_2(u)f\| + \|P_2(u^\perp)f\| = 1.$$

Moreover, by [7, Proposition 1.1(d)] we have

$$\big(\|P_2(u)f\|g + \|P_2(u^\perp)f\|h\big)\,(u + u^\perp) = \|P_2(u)f\| + \|P_2(u^\perp)f\| = 1.$$

Thus,

$$\|P_2(u)f\|g + \|P_2(u^\perp)f\|h \in F_{u+u^\perp}.$$

□

Lemma 3.3. *Let Z be a neutral SFS-space. If ω is a complete geometric tripotent, then for any $u < \omega$ we have:*

$$F_u \oplus_c F_{u^\perp} = \left\{f \in F_{u+u^\perp} :\, f = \|P_2(u)f\|\, g + \left\|P_2(u^\perp)f\right\| h,\, g \in F_u,\, h \in F_{u^\perp}\right\}. \tag{3.2}$$

Proof: Denote the expression on the right-hand side of equality (3.2) by A. Let $\forall f \in F_u \oplus F_{u^\perp}$, i.e. $f = \alpha g + \beta h$, where $\alpha, \beta \geq 0$, $\alpha + \beta = 1$, $g \in F_u$ and $h \in F_{u^\perp}$. Since $g \diamondsuit h$, by [7, Proposition 1.1] and [11, (16)] we have

$$\alpha = f(u) = \|P_2(u)f\|, \quad \beta = f(u^\perp) = \left\|P_2(u^\perp)f\right\|,$$

hence $f \in A$, and therefore $F_u \oplus F_{u^\perp} \subset A$.

Conversely, let $\forall f \in A$, i.e.

$$f = \|P_2(u)f\|\, g + \left\|P_2(u^\perp)f\right\| h, \quad g \in F_u,\, h \in F_{u^\perp}. \tag{3.3}$$

From Lemma 3.2(i) it follows that $f \in co(F_u \cup F_{u^\perp})$. It remains to prove the uniqueness of representation (3.3).

First, consider the case when $\|P_2(u)f\| \neq 0$ and $\left\|P_2(u^\perp)f\right\| \neq 0$. Suppose

$$f = \|P_2(u)f\|\, g + \left\|P_2(u^\perp)f\right\| h = \|P_2(u)f\|\, g' + \left\|P_2(u^\perp)f\right\| h',$$

where $g, g' \in F_u$ and $h, h' \in F_{u^\perp}$. Since $Z_0(u), Z_2(u)$ are subspaces of Z, and $g, g' \in Z_2(u)$, $h, h' \in Z_0(u)$, then by [8, Proposition 1.5] we obtain

$$\|P_2(u)f\|\,(g - g') \diamondsuit \left\|P_2(u^\perp)f\right\|\,(h - h').$$

Hence,

$$\|P_2(u)f\|\,\|g-g'\| + \left\|P_2(u^\perp)f\right\|\,\|h-h'\| = 0.$$

Since $\|P_2(u)f\| \neq 0$ and $\left\|P_2(u^\perp)f\right\| \neq 0$, it follows that $\|g-g'\| = 0$ and $\|h-h'\| = 0$, i.e. $g = g'$ and $h = h'$.

Now consider the cases when $\|P_2(u)f\| = 0$ or $\left\|P_2(u^\perp)f\right\| = 0$. Then $f = \left\|P_2(u^\perp)f\right\| h$ or $f = \|P_2(u)f\|\, g$, respectively. Uniqueness in these cases is obvious. Note that $\|P_2(u)f\|$ and $\left\|P_2(u^\perp)f\right\|$ cannot both be zero simultaneously, since $\|f\| = 1$.

Thus, $F_u \oplus_c F_{u^\perp} = A$. □

Proof of theorem 3.1. From [9, Theorem 2.2] it follows that F_u is a strongly split face, i.e. $F_u \oplus_c F_{u^\perp} = F_\omega$. Hence, from (3.2) it follows (3.1). □

Theorem 3.4. *Let Z be a real neutral SFS-space with a unitary geometric tripotent. Then any symmetric face F_ω corresponding to a maximal geometric tripotent ω admits the following representation:*

$$F_\omega = \left\{f \in Z_1 :\ f = \|P_2(u)f\|\, g + \left\|P_2(u^\perp)f\right\| h,\ g \in F_u,\ h \in F_{u^\perp}\right\}, \tag{3.4}$$

where $u \in G\mathfrak{U}$ and $u < \omega$.

We first prove the following auxiliary lemmas.

Lemma 3.5. *If u is not a maximal geometric tripotent, then there exists a geometric tripotent e such that $u < e$.*

Proof: Since $P_0(u) \neq 0$, there exists a nonzero element $f \in Z_0(u) \cap \partial Z_1$. Then $v(f) \in U_0(u)$ (see [11, §2]), and hence, by [7, Lemma 2.5], it follows that $v(f)\diamondsuit u$ and $v(f) + u \in G\mathfrak{U}$. Therefore, $u < v(f) + u$. □

Lemma 3.6. *Let u, v be mutually orthogonal geometric tripotents. Then $u+v$ is maximal if and only if $u - v$ is maximal.*

Proof: Suppose $u + v$ is maximal. Assume that $u - v$ is not a maximal geometric tripotent. Then, by Lemma 3.5, there exists a geometric tripotent e such that $u - v < e$. Set $\omega = e - (u - v)$. Then, according to [8, Lemma 4.2], we have $\omega\diamondsuit(u - v)$. Consequently, from [7, Corollary 2.6] it follows that $\omega\diamondsuit u$ and $\omega\diamondsuit v$. Hence, by [21, Lemma 3.2(e)] we obtain $\omega\diamondsuit(u + v)$. Therefore, $u + v \leq u + v + \omega$. This contradicts the maximality of $u + v$. □

Lemma 3.7. *If e is a maximal geometric tripotent, then any symmetric face $F_u \subset F_e$ is strongly split.*

Proof: First, we show that $F_e = co\{F_u \cup F_{u^\perp}\}$. For this, it suffices to establish the inclusion $F_e \subseteq co\{F_u \cup F_{u^\perp}\}$. Since e is maximal and $e = u + u^\perp$, it follows from Lemma 3.6 that the geometric tripotent $u - u^\perp$ is maximal, and therefore unitary (see [22, Lemma 4.1]). Hence,

$$Z_1 = co\{F_{u-u^\perp} \cup F_{u^\perp - u}\}.$$

Thus, every element $f \in F_e$ can be written in the form

$$f = tg + (1-t)h$$

for some $g, h \in F_{u-u^\perp}, F_{u^\perp - u}$ respectively, with $0 \leq t \leq 1$.

We now consider the following cases.

Case 1. If $t = 0$, then $f \in F_e \cap F_{u^\perp - u} = F_{u+u^\perp} \cap F_{u^\perp - u} = F_{u^\perp}$.

Case 2. If $t = 1$, then $f \in F_e \cap F_{u-u^\perp} = F_{u+u^\perp} \cap F_{u-u^\perp} = F_u$.

Case 3. If $0 < t < 1$, then applying the geometric tripotent $e = u + u^\perp$ to $f = tg + (1-t)h$, we obtain

$$1 = \langle e, f\rangle = \langle u + u^\perp, tg + (1-t)h\rangle = t\langle u, g\rangle + t\langle u^\perp, g\rangle + (1-t)\langle u, h\rangle + (1-t)\langle u^\perp, h\rangle. \tag{3.5}$$

Since $g \in F_{u-u^\perp}$ and $h \in F_{u^\perp-u}$, we have

$$1 = \langle u - u^\perp, g\rangle = \langle u, g\rangle - \langle u^\perp, g\rangle, \quad 1 = \langle u^\perp - u, h\rangle = \langle u^\perp, h\rangle - \langle u, h\rangle.$$

Thus,

$$\begin{aligned} 1 = t + 1 - t = t(\langle u, g\rangle - \langle u^\perp, g\rangle) + (1-t)(\langle u^\perp, h\rangle - \langle u, h\rangle) = \\ = t\langle u, g\rangle - t\langle u^\perp, g\rangle + (1-t)\langle u^\perp, h\rangle - (1-t)\langle u, h\rangle. \end{aligned} \tag{3.6}$$

From (3.5) and (3.6) it follows that

$$t\langle u, g\rangle + (1-t)\langle u^\perp, h\rangle = 1.$$

Since $\left|\langle u, g\rangle\right| \le 1$ and $\left|\langle u^\perp, h\rangle\right| \le 1$, this equality implies that

$$\langle u, g\rangle = \langle u^\perp, h\rangle = 1.$$

Hence, $g \in F_u$ and $h \in F_{u^\perp}$. Therefore,

$$f = tg + (1-t)h \in co\{F_u \cup F_{u^\perp}\},$$

which proves that $F_e \subseteq co\{F_u \cup F_{u^\perp}\}$.

Thus, $F_e = co\{F_u \cup F_{u^\perp}\}$. Taking into account that $F_u \diamondsuit F_{u^\perp}$, we obtain $F_e = F_u \oplus_c F_{u^\perp}$, i.e. the symmetric face F_u is strongly split. □

Proof of theorem 3.4. Since ω is a maximal geometric tripotent, it follows from Lemma 3.7 that any symmetric face $F_u \subset F_\omega$ is strongly split. Then, from Theorem 3.1 it follows that F_ω admits the representation (3.4). □

From Theorem 3.4 the following statements follow.

Corollary 3.8. *The unit ball admits the following representation:*

$$Z_1 = \left\{ f \in Z : \, f = \|P_2(u) f\| \, g + \left\|P_2\left(u^\perp\right) f\right\| h, \, g \in co\left(F_u \cup F_{-u}\right), \, h \in co\left(F_{u^\perp} \cup F_{-u^\perp}\right) \right\}.$$

Corollary 3.9. *Let ω be a maximal geometric tripotent.*

(i) *For any geometric tripotent $u < \omega$ we have*

$$Z = Z_2(u) \oplus Z_2\left(u^\perp\right), \quad \text{and} \quad P_1(u) = 0.$$

(ii) *Let u_1 and u_2 be mutually orthogonal geometric tripotents such that $u_1, u_2 < \omega$. Then the following equality holds:*

$$P_2(u_1 + u_2) = P_2(u_1) + P_2(u_2).$$

Proof: (i) From the unitarity of ω (see [22, Lemma 4.1]) and the strong splitness of F_u (see Lemma 3.7) we obtain

$$Z = sp\, Z_1 = sp\left\{co\left(F_e \cup F_{-e}\right)\right\} = sp\, F_e = sp\left(F_u \oplus_c F_{u^\perp}\right) = sp\, F_u \oplus sp\, F_{u^\perp}.$$

Since $sp\, F_u \diamondsuit sp\, F_{u^\perp}$, it follows that $\overline{sp}\, F_u \diamondsuit \overline{sp}\, F_{u^\perp}$, and hence

$$Z = \overline{sp} F_u \oplus \overline{sp} F_{u^\perp} = Z_2(u) \oplus Z_2\left(u^\perp\right).$$

This implies that $P_2(u) + P_2\left(u^\perp\right) = I$. Since $P_1(u) P_0(u) = 0$ and $P_2\left(u^\perp\right) = P_0(u) P_2\left(u^\perp\right)$ (see [8, Corollary 3.4]), we obtain $P_1(u) P_2\left(u^\perp\right) = 0$. Therefore,

$$P_1(u) = P_1(u) I = P_1(u)\left[P_2(u) + P_2\left(u^\perp\right)\right] = 0.$$

(ii) From [8, Lemma 1.8] we have

$$P_0(u_1 + u_2) = P_0(u_1) P_0(u_2).$$

Using this equality and taking into account $P_1(u_1) = P_1(u_2) = P_1(u_1+u_2) = 0$, as well as [8, Corollary 3.4], we obtain:

$$\begin{aligned} P_2(u_1+u_2) &= I - P_0(u_1+u_2) = I^2 - P_0(u_1+u_2) \\ &= (P_2(u_1)+P_0(u_1))(P_2(u_2)+P_0(u_2)) - P_0(u_1)P_0(u_2) \\ &= P_2(u_1)+P_2(u_2)+P_0(u_1)P_0(u_2) - P_0(u_1)P_0(u_2) \\ &= P_2(u_1)+P_2(u_2). \end{aligned}$$

□

References

[1] Ayupov S. A., Yadgorov N. Z., Convex spectral sets in finite-dimensional spaces. Izvestiya Akademii Nauk UzSSR, Seriya Fiziko-Matematicheskikh Nauk, (1989), Iss. 3, P. 3–7.

[2] Ayupov S. A., Yadgorov N. Z., Properties of spectral convex sets. Doklady Akademii Nauk UzSSR, (1989), Iss. 7, P. 3–4.

[3] Ayupov S. A., Yadgorov N. Z., Geometry of the state space of modular Jordan algebras. Russian Academy of Sciences. Izvestiya Mathematics, (1994) Vol. 43, Iss. 5, P. 581–592.

[4] Ayupov S. A., Iochum B., Yadgorov N., Symmetry versus facial homogeneity for self-dual cones. Linear Algebra and its Applications, (1990) Vol. 142, P. 83–89.

[5] Alfsen E., Hanche-Olsen H., Shultz F. W., State spaces of C^*-algebras. Acta Mathematica, (1980) Vol. 144, P. 267–305.

[6] Alfsen E., Shultz F. W., State spaces of Jordan algebras. Acta Mathematica, (1978) Vol. 140, P. 155–190.

[7] Friedman Y., Russo B., A geometric spectral theorem. Quarterly Journal of Mathematics, Oxford, (1986) Vol. 37, Iss. 2, P. 263–277.

[8] Friedman Y., Russo B., Affine structure of facially symmetric spaces. Mathematical Proceedings of the Cambridge Philosophical Society, (1989) Vol. 106, P. 107–124.

[9] Ibragimov M. M., Geometric properties of geometric tripotents and split faces in neutral SFS-space. Uzbek Mathematical Journal, (2024) Vol. 68, Iss. 2, P. 76–80.

[10] Friedman Y., Russo B., Classification of atomic facially symmetric spaces. Canadian Journal of Mathematics, (1993) Vol. 45, Iss. 1, P. 33–87.

[11] Friedman Y., Russo B., Geometry of the dual ball of the spin factor. Proceedings of the London Mathematical Society, (1992) Vol. 65, P. 142–174.

[12] Friedman Y., Russo B., Some affine geometric aspects of operator algebras. Pacific Journal of Mathematics, (1989) Vol. 137, P. 123–144.

[13] Friedman Y., Russo B., Structure of the predual of a JBW^*-triple. Journal für die reine und angewandte Mathematik, (1985) Vol. 356, P. 67–89.

[14] Garcia M. C., Palacios M. R., Non-associative Normed Algebras. Vol. 1. Encyclopedia of Mathematics and its Applications, Vol. 154. Cambridge University Press, Cambridge, (2014).

[15] Dang T., Friedman Y., Russo B., Affine geometric proofs of the Banach–Stone theorems of Kadison and Kaup. Rocky Mountain Journal of Mathematics, (1990) Vol. 20, Iss. 2, P. 395–407.

[16] Edwards C. M., Ruttimann G. T., On the facial structure of the unit balls in a JBW^*-triple and its predual. Journal of the London Mathematical Society, (1988) Vol. 38, P. 317–332.

[17] Edwards C. M., Ruttimann G. T., Orthogonal faces of the unit ball in a Banach space. Atti Sem. Mat. Fis. Univ. Modena, (2001) Vol. 49, P. 473–493.

[18] Ibragimov M. M., Geometric properties of neutral facially symmetric spaces. Uzbek Mathematical Journal, (2023) Vol. 67, Iss. 2, P. 60–66.

[19] Ibragimov M. M., Geometric properties of a unit ball of a neutral facially symmetric space. Bulletin of the Institute of Mathematics, (2022) Vol. 5, Iss. 2, P. 51–55.

[20] Friedman Y., Physical Applications of Homogeneous Balls. Springer, New York, (2005).

[21] Ibragimov M. M., Orthoisomorphisms between lattices of geometric tripotents in dual spaces of facially symmetric spaces. Uzbek Mathematical Journal, (2024) Vol. 68, Iss. 3, P. 66–75.

[22] Kudaybergenov K. K., Seypullaev J. X., Description of facially symmetric spaces with unitary tripotents. Siberian Advances in Mathematics, (2020) Vol. 30, Iss. 2, P. 77–83.

Ibragimov M.M.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences, Tashkent, Uzbekistan
e-mail: mukhtar_i@bk.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 111–114
DOI: 10.29229/uzmj.2026-2-13

Integer Powers and Catalan- Mihăilescu Theorem

Kader S.

Abstract. Catalan- Mihăilescu Theorem states that the only case of difference between two consecutive powers equals 1 is $3^2 - 2^3$. This article provides some elementary analyses of integer powers, and proves in that light, Catalan- Mihăilescu Theorem.

Keywords: Integer, power, consecutive, co-prime
MSC (2020): 11D61, 11D41, 11D45

1. Introduction

Catalan's conjecture(or Mihăilescu's theorem) [1, 2] is a theorem in number theory that was conjectured by the Belgian mathematician Eugène Charles Catalan (1814-1894).The integers 2^3 and 3^2 are two powers of natural numbers whose values (8 and 9, respectively) are consecutive. The theorem states that this is the only case of two consecutive powers. That is to say, that the only solution in the natural numbers of the equation:

$$x^m - y^n = 1$$

for $m, n > 1$; $x, y > 0$ is $x = 3, m = 2, y = 2, n = 3$.
Catalan mentioned this conjecture as a problem in the section Théorèmes et problèmes on p. 520 of the first volume of the journal Nouvelles Annales de Mathématiques in 1842. [3] The conjecture was received by the editor of the Journal für die Reine und Angewandte Mathematik as a letter from Catalan, and published in the journal in 1844. [4] Catalan, at that time a teacher at l'_Ecole Polytechnique de Paris, had won his reputation with a solution of a combinatorial problem. The term Catalan number, still in use, refers to that problem. As to the equation, Catalan wrote that he could not prove it completely so far. He never published any serious partial result about it either [1]. The only results of Catalan on this equation are simple observations in his Mélanges Mathématiques,XV, published much later in 1885.[5] Catalan's conjecture was proven by Preda Mihăilescu in April 2002. The proof was published in the Journal für die reine und angewandte Mathematik, 2004. It makes extensive use of the theory of cyclotomic fields and Galois modules[6]. This article provides some elementary analyses of integer powers, and proves in that light, Catalan- Mihăilescu Theorem.

2. Integer Powers and Catalan- Mihăilescu Theorem

Lemma 2.1. *When x, y, m, n are positive integers, $x^m - y^n = 1$ then we have m, n co-primes.*

Proof: Suppose, m, n are not co-primes. Let $m = r\mu, n = r\eta$ when μ, η are positive integers. Then

$$x^m - y^n = x^{r\mu} - y^{r\eta} = (x^\mu)^r - (y^\eta)^r = (x^\mu - y^\eta)((x^\mu)^{r-1} + \ldots (y^\eta)^{r-1}) \neq 1$$

Contradicts our equation. So, m, n are co-primes. □

Lemma 2.2. *When x, y, m, n are positive integers, $x^m - y^n = 1$ then we have x, y co-primes.*

Proof: Suppose, x, y are not co-primes. Let $x = r\alpha, y = r\beta$ when α, β are positive integers, $\alpha \neq \beta$. Then, if $n > m$ we have,

$$x^m - y^n = (r\alpha)^m - (r\beta)^n = r^m(\alpha^m - r^{n-m}\beta^n) \neq 1$$

if $m > n$ we have,

$$x^m - y^n = (r\alpha)^m - (r\beta)^n = r^n(r^{m-n}\alpha^m - \beta^n) \neq 1$$

if $m = n$ we have,

$$x^m - y^n = (r\alpha)^m - (r\beta)^n = (r\alpha)^m - (r\beta)^m = r^m(\alpha^m - \beta^m) \neq 1$$

All contradict our equation. So, x, y are co-primes. □

Lemma 2.3. *When x, y, m, n are positive integers, $x^m - y^n = 1$ then x, y are not both odd.*

Proof: Suppose, x, y are both odd. Then when k is an integer,

$$x^m - y^n = 2k \neq 1$$

Contradicts our equation. So, x, y are not both odd. □

Lemma 2.4. *$\forall u > 1, w \in \mathbb{N}$, $\exists z \in \mathbb{N}$, such that, $u^{2w} = (u+1)z + 1$ where $z = u(z_i u + 1) - 1$ when $0 = z_0 \leq z_i < z$, and z_i is immediate previous value of z . The formula defines parity of z which is odd when u even, even when u odd iterating over $z_0 = 0$.*

Proof: We see,

$2^{2\times1} = 3 \times 1 + 1 = 4$	$3^{2\times1} = 4 \times 2 + 1 = 9$	$4^{2\times1} = 5 \times 3 + 1 = 16$
$2^{2\times2} = 3 \times 5 + 1 = 16$	$3^{2\times2} = 4 \times 20 + 1 = 81$	$5^{2\times1} = 6 \times 4 + 1 = 25$
$2^{2\times3} = 3 \times 21 + 1 = 64$	$3^{2\times3} = 4 \times 182 + 1 = 729$	$6^{2\times1} = 7 \times 5 + 1 = 36$

Here we see the pattern, $u^{2w} = (u+1)z + 1$.
By induction method, the next expression,

$$u^{2(w+1)} = u^{2w+2} = u^2 \times u^{2w} = u^2((u+1)z + 1)$$
$$= (u+1)zu^2 + u^2 = (u+1)zu^2 + u^2 - 1 + 1 = (u+1)(zu^2 + u - 1) + 1.$$

Letting $w + 1 = w_1$ and $zu^2 + u - 1 = u(zu + 1) - 1 = z_1$, the pattern of expression remains the same,

$$u^{2w_1} = (u+1)z_1 + 1$$

So by induction we prove, $\forall u, w \in \mathbb{N}$, $\exists z \in \mathbb{N}$, such that, $u^{2w} = (u+1)z+1$, and also get the formula, $z = u(z_i u+1) - 1$ when $0 = z_0 \leq z_i < z$, and z_i is immediate previous value of z . The formula defines parity of z which is odd when u even, even when u odd iterating over $z_0 = 0$. □

Lemma 2.5. *$\forall u > 1, w \in \mathbb{N}$, $\exists z \in \mathbb{N}$, such that, $u^{2w+1} = (u+1)z - 1$ where $z = u(z_i u - 1) + 1$ when $1 = z_0 \leq z_i < z$ ($z_0 = 0$ gives negative value) , and z_i is immediate previous value of z . The formula defines parity of z which is odd when u even, odd when u odd iterating over $z_0 = 1$.*

Proof: We see,

$2^{2\times1+1} = 3 \times 3 - 1 = 8$	$3^{2\times1+1} = 4 \times 7 - 1 = 27$	$4^{2\times1+1} = 5 \times 13 - 1 = 64$
$2^{2\times2+1} = 3 \times 11 - 1 = 32$	$3^{2\times2+1} = 4 \times 61 - 1 = 243$	$5^{2\times1+1} = 6 \times 21 - 1 = 125$
$2^{2\times3+1} = 3 \times 43 - 1 = 128$	$3^{2\times3+1} = 4 \times 547 - 1 = 2187$	$6^{2\times1+1} = 7 \times 31 - 1 = 216$

Here we see the pattern, $u^{2w+1} = (u+1)z - 1$.
By induction method, the next expression,

$$u^{2(w+1)+1} = u^{2w+3} = u^2 \times u^{2w+1} = u^2((u+1)z - 1)$$
$$= (u+1)zu^2 - u^2 = (u+1)zu^2 - u^2 + 1 - 1 = (u+1)(zu^2 - u + 1) - 1.$$

Letting $w + 1 = w_1$ and $zu^2 - u + 1 = u(zu - 1) + 1 = z_1$, the pattern of expression remains the same,

$$u^{2w_1+1} = (u+1)z_1 - 1$$

So by induction we prove, $\forall u, w \in \mathbb{N}$, $\exists z \in \mathbb{N}$, such that, $u^{2w+1} = (u+1)z-1$, and also get the formula $z = u(z_i u - 1) + 1$ when $1 = z_0 \leq z_i < z$ ($z_0 = 0$ gives negative value), and z_i is immediate previous value of z. The formula defines parity of z which is odd when u even, odd when u odd iterating over $z_0 = 1$. □

Corollary 2.6. *When $u > 1, w, v \in \mathbb{N}$, either $w = v$ or $w \neq v$ we have,*
(i)

$$u^{2w} \equiv 1 \mod (u+1)$$

(ii)

$$u^{2v+1} \equiv -1 \mod (u+1)$$

So,

$$u^{2w} \equiv u^{2v+1} + 2 \mod (u+1)$$
$$u^{2w} + u^{2v+1} \equiv 0 \mod (u+1)$$

Theorem 2.7 (Catalan- Mihăilescu Theorem)**.** *When x, y, m, n are positive integers, for*

$$x^m - y^n = 1 \tag{2.1}$$

Only non-trivial solution is $(x, y, m, n) = (3, 2, 2, 3)$

Proof: (i)Relating lemma 2.4 to Catalan- Mihăilescu equation we have,

$$u^{2w} = (u+1)z + 1 = v^g + 1 \tag{2.2}$$

or subtracting 1 from (2.2) we have,

$$v^g = (u+1)z = u^{2w} - 1 = (u^w - 1)(u^w + 1)$$

We see, $u+1|v^g$. It is not possible to make perfect power by $w \geq 2$, as we have then two shown factors factorizing differently with gcd either 1 or 2, we need perfect power factors such that for the gcd = 1 we have $u^w - 1 = a^g$ and $u^w + 1 = b^g$, two more smaller Catalan- Mihăilescu equations (leading to infinite descent), moreover we then have three consecutive integers, a^g, u^w, b^g being perfect powers, two of the same exponent (for $a, b > 1$, minimum gap $b^g - a^g = (b-a)(b^{g-1} + \dots a^{g-1}) \geq b^{g-1} + \dots a^{g-1} \geq b + a$), which is absurd; only possibility with gcd = 1 is when a factor is non-contributory $u^w - 1 = 1$, and $u = 2$, $w = 1$ making $u+1|v^g$, a trivial $g = 1$ and $v^g = u+1 = 3$, the equation (2.2) becomes a trivial, $2^2 = 3 + 1$; for the gcd = 2 the product of the factors will be of the form of $2a^g \cdot 2b^g = 2^2 a^g b^g$ then $g = 2$ but powers of the same exponent, a^g, b^g are consecutive integers which is absurd unless $g = 1$ trivial, $v = 4ab$.
(ii)Relating lemma 2.5 to Catalan- Mihăilescu equation we have,

$$u^{2w+1} = (u+1)z - 1 = v^g - 1 \tag{2.3}$$

or adding 1 to (2.3) we have,

$$v^g = (u+1)z = u^{2w+1} + 1 = (u+1)(u^{2w} - \dots 1) = (u+1)(u^{2w-1}(u-1)\dots + 1)$$

The second factor, is always odd because it has odd number of terms, the last being 1, signs being alternating, the 1st and last terms have positive signs, for $u > 1$ the second factor is not divisible by u, for $u > 2$ the second factor is not divisible by either u or $u - 1$, only divisible by $u + 1$ when $u = 2$, $u - 1 = 1$, $w = 1$, and $u^{2w-1}(u-1)\dots + 1 = u(u-1) + 1 = u + 1$. It is not possible to make perfect power by $w > 1$, odd exponent $2w + 1 > 3$, as we have then two co-prime factors $(u+1) < (u^3(u-1) + u(u-1) + 1) = (u^4 - u^3 + u^2 - u + 1) \leq (u^{2w} - \dots 1)$; only possibility is $u = 2$, $w = 1$, $(u+1)(u+1) = (u+1)^2 = v^g$, $v = 2 + 1 = 3$, $g = 2$, from equation (2.3) we have, the only non-trivial equation, $3^2 = 2^3 + 1$. □

As before the proof by Mihăilescu, then Catalan conjecture had been verified for very big numbers, for which the number with the bigger base of the consecutives and smaller exponent would be smaller because, by their natural logarithms, $\ln(u+1)^w = w \ln(u+1) < w \ln u + \ln u = \ln u^{w+1}$ so, $(u+1)^w < u^{w+1}$ then for big numbers, it could be, $u^{w+1} - (u+1)^w = 1$ or,$(u+1)^w = u^{w+1} - 1 = (u-1)(u^w + \dots + 1)$ which is not possible as sides have un-shared factors unless $w = 1, u = 2$. Moreover, when $s > t > u \geq 2$ we have, $\log_u \frac{t^s}{s^t} = \log_u t^s - \log_u s^t = s \log_u t - t \log_u s > 0 = \log_u 1$. Only exception, $3 \log_2 2 - 2 \log_2 3$. So, $\frac{t^s}{s^t} > 1$ or, $t^s > s^t$.

3. Conclusion

Catalan's Conjecture in number theory is one of those mathematical problems that are very easy to formulate but extremely hard to solve[1]. This conjecture remained unproven since 1844 till 2002.[1, 6] In this article we provided some elementary analyses of integer powers, and in that light we proved Catalan- Mihăilescu Theorem.

Acknowledgment

Those who allowed me hard-found time to do this work.

References

[1] Metsänkylä T., Catalan's Conjecture: Another Old Diophantine Problem Solved. Bulletin (New Series) of the American Mathematical Society, (2003) Vol. 41, Iss. 1, P. 43–57.

[2] Wikipedia contributors. Catalan's conjecture. Wikipedia, the Free Encyclopedia. URL: `https://en.wikipedia.org/wiki/Catalan's_conjecture`.

[3] Bilu Y. F., Bugeaud Y., Mignotte M., The Problem of Catalan. Springer, (2014).

[4] Catalan E., Note extraite d'une lettre adressée à l'éditeur. J. Reine Angew. Math. (1844) Vol. 27, P. 192.

[5] Ribenboim P., Catalan's Conjecture. Academic Press, Boston, (1994).

[6] Preda M., Primary Cyclotomic Units and a Proof of Catalan's Conjecture. J. Reine Angew. Math. (2004) Vol. 572, P. 167–195.

Syed Abdul Kader,
Department of Endocrinology,
Sher-E-Bangla Medical College,
Barisal, Bangladesh
e-mail: skdr0002@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 115–120
DOI: 10.29229/uzmj.2026-2-14

On the negativity of separable algorithmic representations of fields with additive inversions

Karimova N.

Abstract. It is established that any algorithmic representation of any field in a signature with an additive inverse operation that has a nontrivial enumerable subset is negative. In this case, computably and enumerably generated topologies coincide, and the operations are continuous with respect to the enumerably generated topology.

Keywords: field with additive inversion, algorithmic representation, effectively generated topology, negativity and positivity, continuity.
MSC (2020): 03D45

1. Introduction

Following Yu. L. Ershov ([1]), we give a definition of the enumerated universal algebra of effective signature $\sigma = \langle =, f_0^{m_0}, f_1^{m_1}, \dots \rangle$ (where m_i is the number of arguments of the function into which the functional symbol $f_i^{m_i}$ is interpreted and the mapping $h : n \longmapsto m_n$ is computable).

Definition 1.1. An algorithmic representation (or enumeration) of a universal algebra $\mathfrak{A} = \langle A; g_0, g_1, \dots \rangle$ of signature σ is any mapping $\nu : \omega \longrightarrow A$ of the set of positive integers ω onto the underlying set A of $\mathfrak{A}$ such that the following condition holds:

there exists a two-place computable function G such that for any $n \in \omega$, any $y_1, \dots, y_{m_n}$ the equality $g_n(\nu y_1, \dots, \nu y_{m_n}) = \nu G(n, c^{m_n}(y_1, \dots, y_{m_n}))$ holds.

In other words, given ν-indices of elements from A and the index of the operation g_n, we can effectively find some ν-index of the result of applying this operation to these elements.

If $\nu : \omega \longrightarrow A$ is an enumeration of the algebra $\mathfrak{A}$, then the pair $(\mathfrak{A}, \nu)$ is called an enumerated algebra.

Note that every at most countable algebra of effective signature has an enumeration ([1]).

The kernel of the algorithmic representation ν of the enumerated algebra $(\mathfrak{A}, \nu)$ is the equivalence $\{\langle x, y\rangle | \nu x = \nu y\}$. If ν is a representation, then its kernel is denoted by $ker(\nu)$. An enumeration is called computable (positive, negative) if its kernel is computable (computably enumerable, co-enumerable).

For a fixed system, the classical problem is studying its various algorithmic representations and relations between them (see the work [2] of S.S. Goncharov and Yu.L. Ershov), in particular, the problem of the existence of good representations (primarily computable ones) and their number (including uniqueness, up to computable isomorphism).

On the other hand, within the duality paradigm, one can fix the kernel of a representation and study the general properties of systems that have representations with this kernel. This approach seems appropriate from the point of view of the theory of algorithmic representations of systems in a number of problems, including in theoretical computer science (see the reviews [3, 4]).

Much attention is currently being paid to the study of algorithmic properties of equivalences on the set of positive integers. A rich bibliography within the framework of the above-mentioned paradigm is available in S.S.Goncharov and Yu.L.Ershov in [2]. Closely related to this problematic are questions of the structure of algebras representable over equivalences. If η is an equivalence on the set of positive integers ω, then the universal algebra $\mathfrak{A}$ is called representable over η (or an η-algebra) if there exists an enumeration ν with kernel equal to η (i.e. $\eta = \{\langle x, y\rangle \mid \nu x = \nu y\}$). In other words, an algebra $\mathfrak{A}$ is representable over an equivalence η if there exists a family (not necessarily effective) F of computable functions consistent with η (i.e., for each $f \in F : \forall \overline{x}, \overline{y}[\overline{x} = \overline{y} \pmod{\eta} \Rightarrow f(\overline{x}) = f(\overline{y}) \pmod{\eta}]$) such that $\mathfrak{A}$ is isomorphic to the quotient algebra $\langle \omega/\eta; F\rangle$ of the computable algebra $\langle \omega; F\rangle$ by the congruence η. Moreover, the natural projecting homomorphism $\nu(x) = \{x\}/\eta$ is an enumeration from the computable algebra $\langle \omega; F\rangle$ onto $\mathfrak{A}$. It is appropriate to note that the role and

place of computable algebras in the theory of enumerated algebras are similar to the role and place of absolutely free algebras of suitable rank in the theory of universal algebras, since in both cases there exists a homomorphism from a suitable computable (some free) algebra onto $\mathfrak{A}$. Note that any at most countable universal algebra of countable signature has an enumeration (induced by the Gödel enumeration of the absolutely free algebra of terms in a solvable set of generators, [1]). The case of an ineffective signature (of any algorithmic complexity) does not change the situation with the presence of enumeration, i.e. every at most countable algebra of countable signature has an enumeration ([4]).

The word *equivalence* denotes an equivalence relation on the set of positive integers ω. If η is an equivalence, then the set $\alpha \subseteq \omega$ is called *η-closed* if α is the union of suitable η-classes (i.e. $x \in \alpha \wedge x = y \,(\mathrm{mod}\ \eta) \Rightarrow y \in \alpha$).

If η is an equivalence on ω, then the family of η-closed computable (computably enumerable) sets defines a base of the computable (enumerable) topology $\tau(\eta)_C$ (respectively $\tau(\eta)_{CE}$) on the quotient set ω/η.

In complex self-organizing systems, the problem of recognition is apparently of great importance. If a system is specified by its algorithmic representation, then from a mathematical point of view this means that any pair of distinct elements is separated by algorithmically determined neighborhoods. The most important types of separable enumerations of algebras are effectively separable, i.e., those for which there exist computable (in the sense of [5]) families of separating computably enumerable sets. The foundations of the theory of separably enumerated sets and the connections between separability and effective separability were introduced, developed, and investigated by Yu.L.Ershov (see [5]).

Within the framework of the paradigm of the existence of an effective system of algorithmically generated neighborhoods for an enumerated algebra, sufficient for recognizing the separability of any pair of its elements, it must be recognized that it is precisely the most general concept of a separable enumeration studied by Yu.L.Ershov in [5] in the case of enumerations of universal algebras that can be interpreted as one of the mathematical refinements of the concept of a complex developing system, intensionally defined by a family of representing computable functions together with the kernels of homomorphisms (enumerations) and allowing effective recognition of the difference between its constituent elements by separating them with suitable algorithmically determined neighborhoods.

An equivalence η is called computably separated (separable) if for any $x \neq y \ (\mathrm{mod}\ \eta)$ there exists a computable η-closed set containing x and not containing y (there exists a computably enumerable η-closed set α such that either $x \in \alpha \wedge y \notin \alpha$ or $y \in \alpha \wedge x \notin \alpha$).

Let (N, ν) be an enumerated set. A subset N_0 of N is called *ν-computable (ν-enumerable, ν-co-enumerable)* if the set $\nu^{-1}N_0$ is computable (computably enumerable, co-enumerable). If it is clear from the context which enumeration of a set is meant, then we will call its subsets simply *computable (enumerable, co-enumerable)* without the prefix ν.

Each enumeration ν is uniquely associated with its kernel (enumeration equivalence), i.e. the set $ker(\nu) = \{\langle x, y\rangle \mid \nu x = \nu y\}$ and when speaking about certain properties of an enumeration we often imply the presence of these properties for the kernel. An enumeration is called computably separable (separable) if its kernel is computably separable (separable).

An enumeration ν is called effectively separable if its kernel $ker(\nu)$ is so, i.e. there exists a computable family $\mathfrak{S}$ of computably enumerable $ker(\nu)$-closed sets such that for any $x, y \in \omega$ if $\nu x \neq \nu y$ then there exists $S \in \mathfrak{S}$ such that either $\nu x \in \nu S \wedge \nu y \notin \nu S$ or $\nu y \in \nu S \wedge \nu x \notin \nu S$ ([5]). Such families is called complete.

Thus, if separability ensures T_0-separability of the space generated by ν-enumerable subsets, then effective separability guarantees the existence of a computable enumeration for a suitable complete separating family.

If $(A, \mu), (B, \nu)$ are two indexed algebras, then the homomorphism $\varphi : A \to B$ is called a morphism if it is effective on indexes, i.e. there exists a computable function f such that the diagram $\varphi\mu = \nu f$ is commutative, which is completely natural from the intensional point of view. Therefore, all the homomorphisms we consider are morphisms.

The most important ones in the framework of the structural theory of computably separated (separable) algorithmic representations of universal algebras are negative and effectively separable algebras, since we have

Theorem 1.2. *An algorithmic representation of an universal algebra is computably separable (separable) if and only if it is approximated by negative algebras (effectively separable algebras).*

Proof in [3, 6].

This fact emphasizes the exceptional role of negative algebras in the theory of separable enumerations of universal algebras. Note that over every negative equivalence a finitely generated congruence-simple algebra is definable ([7]). Similarly, over every negative equivalence with infinitely many cosets a dense linear order is negatively definable ([6]). Nothing similar holds for positive equivalences ([4]). The property of algorithmic uniformity inherent in negative algebras allows to solve some fundamental problems in Computer Science ([8]).

Next, unless otherwise stated, by a topological space on the quotient set ω/η by the equivalence η we mean the enumerably generated space $\tau(\eta)_{CE}$. The case of a $\tau(\eta)_C$-space generated by η-closed computable subsets will be specified separately. It is obvious that the topology of $\tau(\eta)_{CE}$ is generally stronger than the topology of $\tau(\eta)_C$, since computability is a special case of computable enumerability.

2. Topological spaces over computably separated equivalences

Let us formulate the strongest separation properties for computably separable enumerations ([9]):

Proposition 2.1. *For an arbitrary equivalence η on ω the following conditions are equivalent:*

(1) η is a computably separated equivalence;

(2) $\tau(\eta)_C$-space is perfectly normal and totally disconnected.

First of all, we note the following important fact concerning computably separated enumerated universal algebras.

Proposition 2.2. *The operations of any enumerated algebra are continuous in a computably separately generated space.*

Proof: Let $(\omega/\eta;\Im)$ be an arbitrary η-algebra (i.e., an algebra representable over η). Note that we do not assume computability of the family $\Im$ of computable functions for which η is a congruence. If f is a computable function of one variable, then there is nothing to prove, since the preimage of a computable set is also computable. Suppose that $f \in \Im$ and the number of arguments n of f is at least 2. Fix a set $\overline{x} = \langle x_1, \ldots, x_n\rangle$. We need to show that for any η-closed computable set Y containing a number $f(x_1, \ldots, x_n)$, there exist η-closed computable sets $X_1 \ni x_1, \ldots, X_n \ni x_n$ such that $f(X_1, \ldots, X_n) \subseteq Y$. Take the complete f-preimage X of Y, i.e. $X = \{\overline{u} \,|\, f(\overline{u}) \in Y\}$. Clearly, X is a non-empty computable set of tuples of length n if Y is computable. Fix any Gödel enumeration of all tuples in ω^n.

We construct the sets $X_1, \ldots, X_n$ and $Z \subseteq X$ in steps:

Step 0. $X_1^0 = \{x_1\}, \ldots, X_n^0 = \{x_n\}$ and $Z^0 = \varnothing$.

Step $s+1$. Let $\overline{z} = \langle z_1, \ldots, z_n\rangle$ be the first tuple from X that does not belong to the set $X_1^s \times \cdots \times X_n^s \cup Z^s$. If $(X_1^s \cup \{z_1\}) \times \cdots \times (X_n^s \cup \{z_n\}) \subseteq X$, then we put $X_1^{s+1} = X_1^s \cup \{z_1\}, \ldots, X_n^{s+1} = X_n^s \cup \{z_n\}$ and $Z^{s+1} = Z^s$ (in this case, it may turn out that $X_i^{s+1} = X_i^s \cup \{z_i\}$ for some $1 \leqslant i \leqslant n$), otherwise $X_1^{s+1} = X_1^s, \ldots, X_n^{s+1} = X_n^s$ and $Z^{s+1} = Z^s \cup \{\overline{z}\}$. End of step $s+1$.

Now define $X_k = \bigcup_{s\in\omega} X_k^s, 1 \leqslant k \leqslant n$ and $Z = \bigcup_{s\in\omega} Z^s$.

By construction, the computable set X splits into two enumerable disjoint parts — $X_1 \times \cdots \times X_n$ and Z, which implies the computability of the direct product. Note that both of these sets are η-closed, and if some set $\langle z_1, \ldots, z_n\rangle$ is distributed over $X_k, 1 \leqslant k \leqslant n$ at some step (i.e. $z_1 \in X_1, \ldots, z_n \in X_n$), then all sets η-equivalent to this set that appear at later steps are distributed in the same way. The correctness of the construction can easily be shown by induction on the steps of the construction. □

The most important open question: *Is it possible to replace computably separated generated spaces with separably generated spaces in the formulation of proposition 2.2?*

Note that the effective separability of an equivalence means that it lies in the class Π_2^0 of the arithmetic hierarchy, although it does not have a clear "coordinatization" in it, since there exist both effectively separable equivalences outside the class Δ_2^0 and Δ_2^0-equivalences that are not effectively separable, but both Σ_1^0-equivalences and Π_1^0-equivalences lie in the class of effectively separable equivalences ([5]).

3. Effectively non-degenerate representations of fields with additive inversions

Based on the above, a mapping $\nu : \omega \longrightarrow F$ onto the underlying set of an at most countable field $\langle F; +, \bullet, - \rangle$ in a ring signature with the operation of additive inversion is called an algorithmic representation of this field if there exist two binary computable functions f, g such that $\nu x + \nu y = \nu f(x, y)$, $\nu x \bullet \nu y = \nu g(x, y)$ for all $x, y \in \omega$ and one unary computable function h with the property $-\nu x = \nu h(x)$ for all $x \in \omega$.

If η is an equivalence on ω, then F is called representable over η (or an η-field) if there exists a computable algebra of ring signature $R = \langle \omega; +, \bullet \rangle$ with computable operations $+, \bullet$ for which η is a congruence (i.e. $x_0 = y_0 \pmod{\eta} \wedge x_1 = y_1 \pmod{\eta} \Rightarrow x_0 + x_1 = y_0 + y_1 \pmod{\eta}$ and $x_0 = y_0 \pmod{\eta} \wedge x_1 = y_1 \pmod{\eta} \Rightarrow x_0 \bullet x_1 = y_0 \bullet y_1 \pmod{\eta}$) such that F is isomorphic to the quotient algebra $\langle \omega/\eta; +, \bullet \rangle$ of the computable algebra $R = \langle \omega; +, \bullet \rangle$ by the congruence η.

Note that the algebra R may not be a ring. In particular, the laws of associativity and commutativity may not hold in R. The only important thing is that there exists a homomorphism from R onto F, which is a morphism that is an algorithmic representation of ν, where $\nu(x) = \{x\}/\eta$ is the natural projective homomorphism.

Definition 3.1. An algorithmic representation ν of a universal algebra A is called effectively non-degenerate if there exists a non-empty ν-enumerable subset $A_0 \subseteq A$.

Recall that a unary termal operation with fixed elements of an algebra as parameters is called a translation.

Definition 3.2. A universal algebra is called translationally complete if every pair of its distinct elements can be translated into any other pair of distinct elements by a suitable translation.

Note that every translationally complete universal algebra is congruence simple. The converse is not true.

Definition 3.3. A universal algebra is called translationally precomplete if there exists a pair of its distinct elements to which any pair of distinct elements can be translated by a suitable translation.

Obviously, every translationally precomplete universal algebra is subdirectly indecomposable. The converse is not true.

Proposition 3.4. *Every T_2-separable representation of a translationally precomplete algebra is negative.*

Proof: Let (A, ν) be a T_2-separable enumerated translationally precomplete algebra and let a pair of distinct elements $\langle a, b \rangle$ of this algebra be translationally reachable from any pair of distinct elements. Then this pair is also contained in any of its nonzero congruences (i.e. A is subdirectly irreducible). Fix ν-indices of these elements, say $\nu m = a, \nu n = b$. By hypothesis, there exist $ker(\nu)$-closed enumerable disjoint sets α and β such that $m \in \alpha$ and $n \in \beta$.

By T_ν we denote the enumerable set of all translations of A given by the corresponding computable representations of the operations of this algebra in the enumeration ν. Then $x \neq y \Leftrightarrow \exists t \in T_\nu (t(x) \in \alpha \wedge t(y) \in \beta)$. □

In [10] it is shown that the following holds:

Proposition 3.5. *There exists a translationally precomplete algebra possessing a T_1-separable non-Hausdorff effectively separable algorithmic representation ν such that the operations of this algebra are continuous in the enumerable $\tau(ker(\nu))_{CE}$-topology.*

Theorem 3.6. *Every skew field with additive inversion is translationally complete.*

Proof: Let two arbitrary distinct elements a, b of the skew field F be given. Let us fix an arbitrary pair of distinct elements c and d. Let us show that there exists a translation $\lambda x[t(x)]$ (the symbol λ serves for the standard λ-notation) of the skew field F that maps the element a to c, and the

element b to d. So let us define the following set of translations $T(a)$, which have a very simple form of linear polynomials: $T(a) = \{\lambda x[f(x-a)+c] | f \in F\}$. Obviously, every translation $\lambda x[t(x)]$ from $T(a)$ maps the element a to c. At the same time, for any $b \neq a$ there is a unique translation $\lambda x[f_0(x-a)+c] \in T(a)$ that takes element b to d. Clearly, this translation will be $\lambda x[f_0(x-a)+c]$ with $f_0 = (d-c)(b-a)^{-1}$. □

Corollary 3.7. *Every field with additive inversion is translation complete.*

Corollary 3.8. *Every effectively non-degenerate representation of any field with additive inversion is negative.*

Corollary 3.9. *The operations of any field with additive inversion are continuous with respect to the natural topology over any of its effectively non-degenerate algorithmic representations.*

If η is an equivalence on ω, then the η-closure of α, denoted by $[\alpha]_\eta$, is the intersection of all η-closed extensions of α.

A set is called η-infinite (η-finite) if its η-closure consists of an infinite (finite) number of η-equivalence classes. Obviously, a necessary (but not sufficient) condition for a set to be η-infinite is that it is infinite.

If η is an equivalence and δ is a set, then we say that an index x is η-rejected by δ if $x \notin [\delta]_\eta$. It is easy to see that for any fixed negative equivalence η the relation "a positive integer x is η-rejected by a finite set δ" is uniformly enumerable over x, δ (this implies an explicit specification of all elements of the set δ, for example, by means of its canonical index). If it is clear from the context what negative enumeration (or its core) is meant, then we will say that x is rejected by this finite set.

Note that the maximal difference in the topologies of $\tau(\eta)_C$ and $\tau(\eta)_{CE}$ is observed, for example, for perfect equivalences (i.e. positive ones that do not have nontrivial subsets closed with respect to this equivalence, [5]). In this case, $\tau(\eta)_C$-space is trivial, and $\tau(\eta)_{CE}$ is discrete.

For negative equivalences, the situation is diametrically opposite, as shown by

Theorem 3.10. *If ν is a negative enumeration, then the topologies $\tau(\eta)_C$ and $\tau(\eta)_{CE}$ coincide.*

Proof: In one direction this is obvious, since every computable set is computably enumerable.

Let α be a ν-closed subset of ω that is open in the $\tau(\nu)_{CE}$-topology.

Without loss of generality we can assume that α is a basic (i.e. enumerable) neighborhood. Let $a \in \alpha$.

To prove the theorem it suffices to construct a partition of ω into ν-closed enumerable disjoint sets β and δ such that $a \in \beta \subseteq \alpha$ and $\beta \cup \delta = \omega$. We will construct them using a step-by-step construction. The finite parts of the sets β and δ accumulated before step s+1 will be denoted by β_s and δ_s, respectively.

Note that the case $\omega \setminus \alpha = \emptyset$ is also taken into account by the proposed construction.

Step 0. Let $\beta_0 = \{a\}$, $\delta_0 = \emptyset$.

Step s+1. Let z be the smallest positive integer from $\omega \setminus (\beta_s \cup \delta_s)$. Simultaneously enumerating the sets $\omega^2 \setminus ker(\nu)$ and α, we wait until at least one of the following conditions is confirmed (it will be shown later that this will definitely happen):

(a) (z is rejected by the set δ_s) $\wedge$ ($z \in \alpha$);

(b) z is rejected by the set β_s.

If (a) is satisfied, then add z to the enumeration β, otherwise to the enumeration δ.

Lemma 3.11. *The following statements are true:*

(1) $\beta \subseteq \alpha$*;*

(2) $\beta \cap \delta = \emptyset$*;*

(3) any element of β is rejected by all elements of δ and vice versa*;*

(4) $\beta \cup \delta = \omega$.

Proof: Items 1 and 2 are obvious. Item 3 follows from the fact that, as can be easily verified by induction on the steps of the construction, each element of β_s is rejected by all elements of δ_s and vice versa.

Let condition (b) not be satisfied, i.e. z is not rejected by the set β_s. This means that z is ν-equivalent to an element of β_s, and since by virtue of item 1 already proved we have $\beta_s \subseteq \beta \subseteq \alpha$ and α is ν-closed by condition, we obtain $z \in \alpha$. Further, since z is ν-equivalent to some element of β_s, then by virtue of item 3 already proved we obtain that z is rejected by the set δ_s.

To prove item 4, it suffices to show that at each step s one of the conditions (a), (b) in the description of the construction is necessarily satisfied and, as a result, the next z will fall into either β or δ.

Let condition (b) not be satisfied, i.e. z is not rejected by the set β_s. This means that z is ν-equivalent to some element of β_s, and since by virtue of item 1 already proved we have $\beta_s \subseteq \beta \subseteq \alpha$ and α is ν-closed by condition, we obtain $z \in \alpha$. Further, since z is ν-equivalent to some element from β_s, then, by virtue of the already proven item 3, we have the fact of rejection of z by the set δ_s. Thus, in the case of failure to satisfy condition (b), condition (a) will be satisfied. □

The closedness of the sets β and δ with respect to the kernel of the enumeration easily follows from item 3 of the proven lemma 3.11. □

References

[1] Ershov Y. L., Decidability Problems and Constructive Models. Nauka, Moscow, (1980).

[2] Goncharov S. S., Ershov Y. L., Constructive Models. Kluwer Academic/Plenum Publishers, New York, (2000).

[3] Kasymov N. K., Recursively separable enumerated algebras. Russian Mathematical Surveys, (1996) Vol. 51, Iss. 3, P. 509–538.

[4] Kasymov N. K., Dadazhanov R. N., Ibragimov F. N., Separable algorithmic representations of classical systems and their applications. Journal of Mathematical Sciences, (2024) Vol. 278, Iss. 3, P. 476–519.

[5] Ershov Y. L., Theory of Numberings. Nauka, Moscow, (1977).

[6] Kasymov N. K., Homomorphisms onto effectively separable algebras. Siberian Mathematical Journal, (2016) Vol. 57, Iss. 1, P. 36–50.

[7] Kasymov N. K., Algebras over negative equivalences. Algebra and Logic, (1994) Vol. 33, Iss. 1, P. 46–48.

[8] Kasymov N. K., Enumerated algebras with uniformly recursive-separable classes. Siberian Mathematical Journal, (1993) Vol. 34, Iss. 5, P. 869–882.

[9] Kasymov N. K., Separation axioms and partitions of the set of natural numbers. Siberian Mathematical Journal, (1993) Vol. 34, Iss. 3, P. 468–471.

[10] Kasymov N. K., Morozov A. S., Khodzhamuratova I. A., T_1-Separable numberings of subdirectly indecomposable algebras. Algebra and Logic, (2021) Vol. 60, Iss. 4, P. 263–278.

Karimova N. R.,
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: nodirakarimova@bk.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 121–135
DOI: 10.29229/uzmj.2026-2-15

Frames and operators on quaternionic Hilbert spaces

Khachiaa N.

Abstract. The aim of this work is to study frame theory in quaternionic Hilbert spaces. We provide a characterization of frames in these spaces through the associated operators. Additionally, we examine frames of the form $\{Lu_i\}_{i\in I}$, where L is a right $\mathbb{H}$-linear bounded operator and $\{u_i\}_{i\in I}$ is a frame.

Keywords: Frame theory, Operator theory, quaternionic Hilbert spaces
MSC (2020): 42C15, 42C40, 42A38

1. Introduction and preliminaries

Frames in quaternionic Hilbert spaces provide a robust framework for analyzing and reconstructing signals within a higher-dimensional space. As an extension of the classical frame theory, these structures facilitate efficient data representation and processing. The unique properties of quaternionic spaces offer new avenues for exploration, particularly in applications such as signal processing and communications. This work aims to investigate the theoretical foundations of frames in quaternionic Hilbert spaces and their practical implications. The quaternionic field is an extension of the real and complex number systems, consisting of numbers known as quaternions. Quaternions are used to represent three-dimensional rotations and orientations, making them invaluable in computer graphics and robotics. Unlike real and complex numbers, quaternion multiplication is non-commutative, which adds complexity to their algebraic structure. Quaternions also provide a more efficient way to perform calculations in three-dimensional space, enhancing applications in physics and engineering.

Definition 1.1 (The field of quaternions)**.** The non-commutative field of quaternions $\mathbb{H}$ is a four-dimensional real algebra with unity. In $\mathbb{H}$, 0 denotes the null element and 1 denotes the identity with respect to multiplication. It also includes three so-called imaginary units, denoted by i, j, k. i.e.,

$$\mathbb{H} = \{a_0 + a_1 i + a_2 j + a_3 k : \ a_0, a_1, a_2, a_3 \in \mathbb{R}\},$$

where $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$ and $ki = -ik = j$. For each quaternion $q = a_0 + a_1 i + a_2 j + a_3 k$, we define the conjugate of q denoted by $\overline{q} = a_0 - a_1 i - a_2 j - a_2 k \in \mathbb{H}$ and the module of q denoted by $|q|$ as

$$|q| = (\overline{q}q)^{\frac{1}{2}} = (q\overline{q})^{\frac{1}{2}} = \sqrt{a_0^2 + a_1^2 + a_2^2 + a_3^2}.$$

For every $q \in \mathbb{H}$, $q^{-1} = \dfrac{\overline{q}}{|q|^2}$.

Because $\mathbb{H}$ is non-commutative, one has to specify whether scalar multiplication is taken on the left or on the right when dealing with quaternionic vector spaces. Throughout this work, we adopt the convention of right scalar multiplication. With slight modifications, the corresponding results remain valid for the left scalar multiplication setting as well.

Definition 1.2 (Right quaternionic vector space)**.** A right quaternionic vector space V is a linear vector space under right scalar multiplication over the field of quaternions $\mathbb{H}$, i.e., the right scalar multiplication

$$\begin{array}{rcl} V \times \mathbb{H} & \to & V \\ (v, q) & \mapsto & v.q, \end{array}$$

satisfies the following for all $u, v \in V$ and $q, p \in \mathbb{H}$:

(1) $(v + u).q = v.q + u.q$,

(2) $v.(p + q) = v.p + v.q$,

(3) $v.(pq) = (v.p).q$.

Instead of $v.q$, we often use the notation vq.

Definition 1.3 (Right quaterninoic pre-Hilbert space)**.** A right quaternionic pre-Hilbert space $\mathcal{H}$, is a right quaternionic vector space equipped with the binary mapping $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{H}$ (called the Hermitian quaternionic inner product) which satisfies the following properties:

(*a*) $\langle v_1, v_2 \rangle = \overline{\langle v_2, v_1 \rangle}$ for all $v_1, v_2 \in \mathcal{H}$,

(*b*) $\langle v, v \rangle > 0$ if $v \neq 0$,

(*c*) $\langle v, v_1 + v_2 \rangle = \langle v, v_1 \rangle + \langle v, v_2 \rangle$ for all $v, v_1, v_2 \in \mathcal{H}$,

(*d*) $\langle v, uq \rangle = \langle v, u \rangle q$ for all $v, u \in \mathcal{H}$ and $q \in \mathbb{H}$.

In view of Definition 3 , a right pre-Hilbert space $\mathcal{H}$ also has the property:

(i) $\langle vq, u \rangle = \overline{q} \langle v, u \rangle$ for all $v, u \in \mathcal{H}$ and $q \in \mathbb{H}$.

Let $\mathcal{H}$ be a right quaternionic pre-Hilbert space with the Hermitian inner product $\langle \cdot, \cdot \rangle$. Define the quaternionic norm $\| \cdot \| : \mathcal{H} \to \mathbb{R}^+$ on $\mathcal{H}$ by

$$\|u\| = \sqrt{\langle u, u \rangle}, \quad u \in \mathcal{H},$$

which satisfies the following properties:

(1) $\|uq\| = \|u\||q|$, for all $u \in \mathcal{H}$ and $q \in \mathbb{H}$,

(2) $\|u + v\| \leq \|u\| + \|v\|$, for all $u, v \in \mathcal{H}$.

(3) $\|u\| = 0 \Longleftrightarrow u = 0$, for all $u \in \mathcal{H}$.

Definition 1.4 (Right quaternionic Hilbert space)**.** A right quaternionic pre-Hilbert space is called a right quaternionic Hilbert space if it is complete with respect to the quaternionic norm.

Example 1.5. Let I be a countable set . Define

$$\ell^2(I, \mathbb{H}) := \left\{ \{q_i\}_{i \in I} \subset \mathbb{H} : \sum_{i \in I} |q_i|^2 < \infty \right\}.$$

$\ell^2(\mathbb{H})$ under right multiplication by quaternionic scalars together with the quaternionic inner product defined as: $\langle p, q \rangle := \sum_{i \in I} \overline{p_i} q_i$ for $p = \{p_i\}_{i \in I}$ and $q = \{q_i\}_{i \in I} \in \ell^2(\mathbb{H})$, is a right quaternionic Hilbert space.

Theorem 1.6 (The Cauchy-Schwarz inequality)**.** *[1] If $\mathcal{H}$ is a right quaternionic Hilbert space, then for all $u, v \in \mathcal{H}$,*

$$|\langle u, v \rangle| \leq \|u\| \|v\|.$$

Definition 1.7 (orthogonality)**.** Let $\mathcal{H}$ be a right quaternionic Hilbert space and A be a subset of $\mathcal{H}$. Then, define the set:

- $A^\perp = \{v \in \mathcal{H} : \langle v, u \rangle = 0 \ \forall \ u \in A\}$;
- $\langle A \rangle$ as the right quaternionic vector subspace of $\mathcal{H}$ consisting of all finite right $\mathbb{H}$-linear combinations of elements of A.

Proposition 1.8. *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and A be a subset of $\mathcal{H}$. Then,*

(1) $A^\perp = \langle A \rangle^\perp = \overline{\langle A \rangle}^\perp = \overline{\langle A \rangle^\perp}$.

(2) $(A^\perp)^\perp = \overline{\langle A \rangle}$.

(3) $\overline{A} \oplus A^\perp = \mathcal{H}$.

Theorem 1.9. *[1] Let $\mathcal{H}$ be a quaternionic Hilbert space and let N be a subset of $\mathcal{H}$ such that, for $z, z' \in N$, we have $\langle z, z' \rangle = 0$ if $z \neq z'$ and $\langle z, z \rangle = 1$. Then, the following conditions are equivalent:*

(a) For every $u, v \in \mathcal{H}$, the series $\sum_{z \in N} \langle u, z \rangle \langle z, v \rangle$ converges absolutely and $\langle u, v \rangle = \sum_{z \in N} \langle u, z \rangle \langle z, v \rangle$;

(b) For every $u \in \mathcal{H}$, $\|u\|^2 = \sum_{z \in N} |\langle z, u \rangle|^2$;

(c) $N^{\perp} = \{0\}$;

(d) $\langle N \rangle$ is dense in $\mathcal{H}$.

Definition 1.10. A subset N of $\mathcal{H}$ that satisfies one of the statements in Theorem 2 is called Hilbert basis or orthonormal basis for $\mathcal{H}$.

Theorem 1.11. *[1] Every quaternionic Hilbert space has a Hilbert basis.*

Definition 1.12 (Frames)**.** [2] Let $\{u_i\}_{i \in I}$ be a sequence in a right quaternionic Hilbert space $\mathcal{H}$. $\{u_i\}_{i \in I}$ is said to be Frame for $\mathcal{H}$ if there exist $0 < A \leq B < \infty$ such that for all $u \in \mathcal{H}$, the following inequality holds:

$$A\|u\|^2 \leq \sum_{i \in I} |\langle u_i, u \rangle|^2 \leq B\|u\|^2.$$

(1) If only the upper inequality holds, $\{u_i\}_{i \in I}$ is called a Bessel sequence for $\mathcal{H}$.

(2) If $A = B = 1$, $\{u_i\}_{i \in I}$ is called a Parseval frame for $\mathcal{H}$.

Remark 1.13. Since the terms in the summation in the definition of a frame are non-negative, note that the order of the vectors in a frame is not relevant. That is, if we change the order of the vectors in a frame, the sequence still remains a frame. In other words, if $\{u_i\}_{i \in I}$ is a frame and $\sigma : I \to I$ is a permutation, then $\{u_{\sigma(i)}\}_{i \in I}$ is also a frame.

Example 1.14. Every Hilbert basis for a separable right quaternionic Hilbert space $\mathcal{H}$ is a Parseval frame for $\mathcal{H}$.

Example 1.15. Let $u_1 := (i, 0)$, $u_2 := (0, j)$ and $u_3 := (i, k)$ be three vectors in $\mathbb{H}^2$. Then $\{u_1, u_2, u_3\}$ is a frame for $\mathbb{H}^2$. Indeed, for $u = (v, w) \in \mathbb{H}^2$, we have by the Cauchy-Shwarz inequality 1.6,

$$\sum_{i=1}^{3} \left| \langle u_i, u \rangle_{\mathbb{H}^2} \right|^2 \leq \sum_{i=1}^{3} \|u_i\|_{\mathbb{H}^2}^2 \, \|u\|_{\mathbb{H}^2}^2 = 4\|u\|_{\mathbb{H}^2}^2.$$

Thus, $\{u_1, u_2, u_3\}$ is a Bessel sequence in $\mathbb{H}^2$ with Bessel bound 4.

On the other hand, we have

$$\sum_{i=1}^{3} \left| \langle u_i, u \rangle_{\mathbb{H}^2} \right|^2 = |-iv|^2 + |-jw|^2 + |-iv - kw|^2 \geq |v|^2 + |w|^2 = \|u\|_{\mathbb{H}^2}^2.$$

Hence, $\{u_1, u_2, u_3\}$ is a frame for $\mathbb{H}^2$ with frame bounds 1 and 4.

We can generalize Example 1.15 and state the following result in finite-dimensional right quaternionic Hilbert spaces.

Example 1.16. Let $\mathcal{H}$ be a finite-dimensional right quaternionic Hilbert space and let $\{u_i\}_{i \in I} \subset \mathcal{H}$, where I is a finite set. Then $\{u_i\}_{i \in I}$ is a frame for $\mathcal{H}$ if and only if it spans $\mathcal{H}$. Indeed, Proposition 3.1 shows that every finite frame spans $\mathcal{H}$. Conversely, for all $u \in \mathcal{H}$, we have by Cauchy-Shwarz inequality 1.6,

$$\sum_{i \in I} |\langle u, u_i \rangle|^2 \leq \left(\sum_{i \in I} \|u_i\|^2 \right) \|u\|^2.$$

By setting $B := \sum_{i \in I} \|u_i\|^2$, it follows that $\{u_i\}_{i \in I}$ is a Bessel sequence for $\mathcal{H}$.

Consider the map

$$\begin{array}{rcl} \phi : \mathcal{S} & \longrightarrow & \mathbb{R} \\ u & \mapsto & \sum_{i\in I} |\langle u, u_i\rangle|^2, \end{array}$$

where $\mathcal{S}$ denotes the unit sphere of $\mathcal{H}$. It is clear that ϕ is continuous on the compact set $\mathcal{S}$. Hence, there exists a vector $v \in \mathcal{S}$ such that

$$\sum_{i\in I} |\langle v, u_i\rangle|^2 \leq \sum_{i\in I} |\langle u, u_i\rangle|^2, \qquad \text{for all } u \in \mathcal{S}.$$

Therefore, for any nonzero $u \in \mathcal{H}$, since $\dfrac{u}{\|u\|} \in \mathcal{S}$, we obtain

$$\left(\sum_{i\in I} |\langle v, u_i\rangle|^2\right) \|u\|^2 \leq \sum_{i\in I} |\langle u, u_i\rangle|^2.$$

The case $u = 0$ is trivial. By setting

$$A := \sum_{i\in I} |\langle v, u_i\rangle|^2,$$

we see that $A > 0$ since $\{u_i\}_{i\in I}$ is a spanning family for $\mathcal{H}$. This completes the proof.

This paper is organized as follows. In Section 2, we present some important results from operator theory in the quaternionic setting, which will be used in our study. In Section 3, we explore the reconstruction formula, a fundamental result in frame theory, and define the frame coefficients associated with an arbitrary vector, highlighting their particular features. We also make explicit the optimal bounds of a frame in terms of the associated operators. In Section 4, we characterize sequences of vectors as frames via their associated operators, and investigate the action of bounded operators on frames, providing a characterization of those that transform one frame into another. We further examine when two frames are related by a bounded operator. At the end of this section, we discuss the correspondence between frames and bounded self-adjoint, positive, and invertible operators. Finally, the correspondence between frames and bounded operators into $\ell^2(I, \mathbb{H})$ is also addressed.

2. Auxiliary results from operator theory in quaternionic setting

In this section, we will present some interesting results on operator theory in quaternionic Hilbert spaces, which will be utilized in our study. The properties of the associated operators of a frame will also be provided. Note that throughout this section, and in all others, we use a single notation for norms and a single notation for inner products, whenever no confusion arises.

Definition 2.1 (Right $\mathbb{H}$-linear operator)**.** [1] Let $\mathcal{H}$ and $\mathcal{K}$ be two right quaternionic Hilbert spaces. Let $L : \mathcal{H} \to \mathcal{K}$ be a map.

(1) L is said to be right $\mathbb{H}$-linear operator if $L(uq + vp) = L(u)q + L(v)p$ for all $u, v \in \mathcal{H}$ and $p, q \in \mathbb{H}$.

(2) If L is a right $\mathbb{H}$-linear operator. L is continuous if and only if L is bounded; i.e., there exists $M > 0$ such that for all $u \in \mathcal{H}$,

$$\|Lu\| \leq M\|u\|.$$

We denote $\mathbb{B}(\mathcal{H}, \mathcal{K})$ the set of all right $\mathbb{H}$-linear bounded operators from $\mathcal{H}$ to $\mathcal{K}$, and if $\mathcal{H} = \mathcal{K}$, we denote $\mathbb{B}(\mathcal{H})$ instead of $\mathbb{B}(\mathcal{H}, \mathcal{H})$.

(3) If L is a right $\mathbb{H}$-linear bounded operator, we define the norm of L as:

$$\|L\| = \sup_{\|u\|=1} \|Lu\| = \inf\{M > 0 : \ \|Lu\| \leq M\|u\|, \ \forall u \in \mathcal{H}\}.$$

And we have for all $L, M \in \mathbb{B}(\mathcal{H}), \|L + M\| \leq \|L\| + \|M\|$ and $\|MN\| \leq \|L\|\|M\|$.

Theorem 2.2 (Quaternionic representation Riesz' theorem)**.** *[1] If $\mathcal{H}$ is a right quaternionic Hilbert space, the map*

$$v \in \mathcal{H} \mapsto \langle v \mid \cdot \rangle \in \mathcal{H}'$$

is well-posed and defines a conjugate-$\mathbb{H}$-linear isomorphism.

Theorem 2.3 (The uniform boundedness principle)**.** *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and F be any subset F of $\mathbb{B}(\mathcal{H})$, if*

$$\sup_{L \in F} \|Lv\| < +\infty \quad \textit{for every } v \in \mathcal{H},$$

then:

$$\sup_{L \in F} \|L\| < +\infty.$$

Theorem 2.4 (The open map theorem)**.** *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space. If $L \in \mathbb{B}(\mathcal{H})$ is surjective, then L is open. In particular, if L is bijective, then $L^{-1} \in \mathbb{B}(\mathcal{H})$.*

Theorem 2.5 (The closed graph theorem)**.** *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and let $L : \mathcal{H} \to \mathcal{H}$ be a right $\mathbb{H}$-linear operator. If $Graph(L)$ is closed, then $L \in \mathbb{B}(\mathcal{H})$.*

Definition 2.6 (the adjoint operator)**.** [1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and $L \in \mathbb{B}(\mathcal{H})$. The adjoint operator of L, denoted L^*, is the unique operator in $\mathbb{B}(\mathcal{H})$ satisfying for all $u, v \in \mathcal{H}$:

$$\langle Lu \mid v \rangle = \langle u \mid L^* v \rangle.$$

Definition 2.7. [1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and $L \in \mathbb{B}(\mathcal{H})$.

(1) L is called normal if $LL^* = L^*L$.

(2) L is called self-adjoint if $L = L^*$.

(3) L is called isometric if $\|Lu\| = \|u\|$ for all $u \in \mathcal{H}$.

(4) L is called unitary if $LL^* = L^*L = I$, where I is the identity opertaor of $\mathbb{B}(\mathcal{H})$. An operator is a unitary if and only if it is an isometric surjective operator.

(5) L is called positive, and we write $L \geqslant 0$, if $\langle Lu \mid u \rangle \geqslant 0$ for all $u \in \mathcal{H}$.

For a linear operator L, we denote by $R(L)$ its range and by $\ker(L)$ its kernel.

Proposition 2.8. *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and $L \in \mathbb{B}(\mathcal{H})$. Then:*

(1) $R(L)^{\perp} = \ker(L^*)$.

(2) $\overline{R(L^*)} = \ker(L)^{\perp}$.

Theorem 2.9 (Square root of an operator)**.** *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and let $L \in \mathbb{B}(\mathcal{H})$. If $L \geq 0$, then there exists a unique operator in $\mathbb{B}(\mathcal{H})$, indicated by $\sqrt{L}$, such that $\sqrt{L} \geq 0$ and*

$$\sqrt{L}\sqrt{L} = L.$$

Furthermore, it turns out that $\sqrt{L}$ commutes with every operator that commutes with L.

Theorem 2.10 (The closed range Theorem)**.** *[1] Let $\mathcal{H}$ be a right quaternionic Hilbert space and let $L \in \mathbb{B}(\mathcal{H})$. If $R(L)$ is closed, then $R(L^*)$ is also closed.*

Proposition 2.11. *Let $\mathcal{H}$ be a right quaternionic Hilbert space and let $L \in \mathbb{B}(\mathcal{H})$. If L is bounded below, i.e. there exists $M > 0$ such that for all $u \in \mathcal{H}$, $M\|u\| \leq \|Lu\|$, and L^* is injective, then L is invertible.*

Proof: Since L is bounded below, then it is injective. Assume that L^* is injective, so $\ker(L^*) = \{0\}$, hence $\overline{R(L)} = \mathcal{H}$. Let us show that $R(L)$ is closed in $\mathcal{H}$. For this, let $y \in \overline{R(L)}$, so there exists a sequence $\{y_n\}_{n \geq 1}$ in $R(L)$ such that $\lim_{n \to \infty} y_n = y$. For each $n \geq 1$, there exists $x_n \in \mathcal{H}$ such that $y_n = L(x_n)$. We have:

$$m\|x_n - x_m\| \leq \|L(x_n) - L(x_m)\| = \|y_n - y_m\|, \quad \forall n, m \geq 1.$$

This implies that $\{x_n\}_{n\geq 1}$ is a Cauchy sequence in $\mathcal{H}$, hence convergent, and let x be its limit. Since L is bounded, we have $y = \lim_{n\to\infty} y_n = \lim_{n\to\infty} L(x_n) = L(x)$. Thus, $y \in R(L)$. Therefore, $R(L)$ is closed, so $R(L) = \mathcal{H}$. Thus, L is surjective and therefore invertible. □

Proposition 2.12. *Let $\mathcal{H}$ be a right quaternionic Hilbert space and let $L \in \mathbb{B}(\mathcal{H})$ be normal. Then, the following statements are equivalent:*

(1) *L is invertible.*

(2) *L is bounded below.*

Proof: It is clear that 1. implies 2. by taking $M = \dfrac{1}{\|L^{-1}\|}$ as the lower bound. Conversely, we use the fact that $\|Lu\| = \|L^*u\|$, $\forall u \in \mathcal{H}$. Then, we apply Proposition 2.11. □

Proposition 2.13. *Let $\mathcal{H}, \mathcal{K}$ be two right quaternionic Hilbert spaces and let $L \in \mathbb{B}(\mathcal{K}, \mathcal{H})$. Then, the following statements are equivalent:*

(1) *L is injective with closed range.*

(2) *L is bounded below.*

(3) *L^* is surjective.*

Proof: Proposition 2.8 and Theorem 2.10 together prove the equivalence (1) $\iff$ (3). Assume that L is injective with closed range, then the operator $L : \mathcal{H} \to R(L)$ is invertible, and since $R(L)$ is closed, it follows, by the open map Theorem 2.4, that $L^{-1} : R(L) \to \mathcal{H}$ is bounded. Now, assume for contradiction that L is not bounded below. Then, for every $n \geq 1$, there exists $u_n \in \mathcal{K}$ with $\|u_n\| = 1$ such that $\dfrac{1}{n} \geq \|L(u_n)\|$, hence

$$Lu_n \xrightarrow[n\to+\infty]{} 0,$$

which implies

$$u_n = L^{-1}(L(u_n)) \xrightarrow[n\to+\infty]{} 0.$$

This is a contradiction since $\|u_n\| = 1$, $\forall n \geq 1$. Conversely, assume that L is bounded below, thus, there exists a strictly positive constant M such that for all $x \in \mathcal{K}$, $M\|x\| \leq \|Lx\|$. It is clear that L is injective. Let $\{x_n\}_{n\geq 1} \subset \mathcal{K}$ be such that

$$Lx_n \xrightarrow[n\to+\infty]{} y \in \mathcal{H}$$

and let's show that $y \in R(L)$. We have

$$\alpha\|x_n - x_m\| \leq \|L(x_n - x_m)\| = \|L(x_n) - L(x_m)\|,$$

thus $\{x_n\}_{n\geq 1}$ is a Cauchy sequence in $\mathcal{K}$, which means it converges to some $x \in \mathcal{K}$. Since L is bounded, $Lx_n \to L(x)$, and by the uniqueness of limits, we conclude that $y = L(x)$ and therefore $y \in R(L)$. Hence, $R(L)$ is closed. □

Let $L \in \mathbb{B}(\mathcal{K}, \mathcal{H})$ be a right $\mathbb{H}$-linear bounded operator with closed range. The restriction of L on $\ker(L)^\perp$, denoted $L_{|\ker(L)^\perp}$, is injective. Indeed, if $x \in \ker(L)^\perp$, we have

$$L_{|\ker(L)^\perp}(x) = 0 \implies x \in \ker(L)^\perp \cap \ker(L) = \{0\} \implies x = 0.$$

On the other hand, we have $L_{|\ker(L)^\perp}(\ker(L)^\perp) = L(\mathcal{K}) = R(L)$. Therefore:

$$L_{|\ker(L)^\perp} : \ker(L)^\perp \to R(L)$$

is invertible. And since $\ker(L)^\perp$ and $R(L)$ are closed, it follows, by the open map Theorem 2.4, that:

$$(L_{|\ker(L)^\perp})^{-1} : R(L) \to \ker(L)^\perp$$

is bounded.

Definition 2.14 (Pseudo-inverse of an operator with closed range)**.** Let $\mathcal{H}$ and $\mathcal{K}$ be two right quaternionic Hilbert spaces and $L \in \mathbb{B}(\mathcal{K}, \mathcal{H})$ be with closed range. The pseudo-inverse of L, denoted by $L^\dagger$, is the right $\mathbb{H}$-linear bounded operator extending $(L_{|ker(L)^\perp})^{-1} : R(L) \to \ker(L)^\perp$ to $\mathcal{H}$ by the property $ker(L^\dagger) = R(L^\perp)$, i.e.,

$$L^\dagger = (L_{|\ker(L)^\perp})^{-1} P_{R(L)},$$

where $P_{R(L)}$ is orthogonoal projection onto $R(L)$.

Theorem 2.15. *Let $L \in \mathbb{B}(\mathcal{K}, \mathcal{H})$ be with closed range and let $v \in R(L)$. The equation*

$$Lx = v$$

admits a unique solution of minimal norm. This solution is exactly $L^\dagger(v)$.

Proof: Let us first show that $L^\dagger(v)$ is a solution. We have $v \in R(L)$, thus $L^\dagger(v) = (L_{|\ker(L)^\perp})^{-1}(v)$, which gives us $L(L^\dagger(v)) = v$. Let $g \in \mathcal{K}$ be a solution to this equation. There exist $g_1 \in \ker(L)$ and $g_2 \in \ker(L)^\perp$ such that $g = g_1 + g_2$. Let us show that $g_2 = L^\dagger(v)$: We have $L(g) = L(g_2)$, and since g is a solution to the equation, we get $L(g_2) = v$. Because g_2 and $L^\dagger(v)$ belong to $\ker(L)^\perp$ and $L_{|\ker(L)^\perp}$ is invertible, it follows that $g_2 = L^\dagger(v)$. Thus, we have $g = g_1 + L^\dagger(v)$. Consequently,

$$\|g\|^2 = \|g_1\|^2 + \|L^\dagger(v)\|^2,$$

which implies that:

$$\|g\| \geq \|L^\dagger(v)\|.$$

Furthermore, we have:

$$\|g\| = \|L^\dagger(v)\| \iff \|g_1\| = 0 \iff g_1 = 0 \iff g = L^\dagger(v).$$

□

Definition 2.16. [2] Let $\{u_i\}_{i\in I}$ be a Bessel sequence for $\mathcal{H}$.

(1) The pre-frame operator of $\{u_i\}_{i\in I}$ is the right $\mathbb{H}$-linear bounded operator denoted by T and defined as follows:

$$\begin{array}{rcl} T : \ell^2(I, \mathbb{H}) & \to & \mathcal{H} \\ q := \{q_i\}_{i\in I} & \mapsto & \sum_{i\in I} u_i q_i. \end{array}$$

(2) The transform operator of $\{u_i\}_{i\in I}$, denoted by θ, is the adjoint of its pre-frame operator. explicitly θ is defined as follows:

$$\begin{array}{rcl} \theta : \mathcal{H} & \to & \ell^2(I, \mathbb{H}) \\ u & \mapsto & \{\langle u_i, u\rangle\}_{i\in I}. \end{array}$$

(3) The frame operator of $\{u_i\}_{i\in I}$, denoted by S, is the composite of T and θ. explicitly, S is defined as follows:

$$\begin{array}{rcl} S : \mathcal{H} & \to & \mathcal{H} \\ u & \mapsto & \sum_{i\in I} u_i \langle u_i, u\rangle. \end{array}$$

Proposition 2.17. *[2] Let $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$. Then:*

(1) *θ is a right $\mathbb{H}$-linear bounded injective operator with closed range.*

(2) *T is a right $\mathbb{H}$-linear bounded surjective operator.*

(3) *S is a right $\mathbb{H}$-linear bounded, positive, and invertible operator.*

Remark 2.18. Let $\{u_i\}_{i\in I}$ be a Bessel sequence for $\mathcal{H}$. Then, for all $u \in \mathcal{H}$, we have:

$$\sum_{i\in I} |\langle u_i, u\rangle|^2 = \langle Su, u\rangle = \|T^* u\|^2 = \|\theta u\|^2.$$

In all what follows, T, θ, and S are reserved to denote the operators defined in the above definition.

3. Frame coefficients and optimal frame bounds

In all this section $\mathcal{H}$ is a separable right quaternionic Hilbert space, $\mathbb{B}(\mathcal{H})$ is the set of all right $\mathbb{H}$-linear bounded operators on $\mathcal{H}$, and I is a countable set.

In what follows, we show that any vector in $\mathcal{H}$ can be expressed (in a generally non-unique way) in terms of the elements of a given frame and that it admits a natural representation with an interesting characteristic.

Proposition 3.1. *Let $\{u_i\}_{i\in I}$ be a frame of $\mathcal{H}$. Then:*

$$\forall u \in \mathcal{H},\ u = \sum_{i\in I} u_i \langle u_i, S^{-1}u\rangle = \sum_{i\in I} S^{-1}u_i \langle u_i, u\rangle.$$

$\rightarrow$ *The coefficients $\{\langle u_i, S^{-1}u\}_{i\in I}$ are called the frame coefficients for u.*

$\rightarrow$ *The expression $u = \sum_{i\in I} u_i \langle u_i, S^{-1}u\rangle$ is often called the natural representation of u.*

Proof: For $u \in \mathcal{H}$, we have: $u = SS^{-1}u = \sum_{i\in I} u_i \langle u_i, S^{-1}u\rangle$, and on the other hand, we have, $u = S^{-1}Su = S^{-1}\left(\sum_{i\in I} u_i \langle u_i, u\rangle\right) = \sum_{i\in I} S^{-1}u_i \langle u_i, u\rangle$. □

Remark 3.2. In general, the decomposition of a vector in a right quaternionic Hilbert space with respect to a frame is not unique, as shown in the following example: Let $(v_i)_{i\geq 1}$ be a Hilbert basis of $\mathcal{H}$. Then the sequence $(v_1, v_1, v_2, v_3, v_4, \ldots)$ is a frame of $\mathcal{H}$. If we set $u = 2v_1$, we can see that u can also be expressed as $u = v_1 + v_1$, which means that $(1, 1, 0, 0, \ldots)$ and $(2, 0, 0, \ldots)$ are two different representations of the same vector u in the frame $(v_1, v_1, v_2, v_3, v_4, \ldots)$.

The following theorem shows the particularity of the frame coefficients.

Theorem 3.3. *Let $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$ and let $u \in \mathcal{H}$ such that $u = \sum_{i\in I} u_i q_i$ where $\{q_i\}_{i\in I} \in \ell^2(I, \mathbb{H})$. Then:*

$$\sum_{i\in I} |q_i|^2 = \sum_{i\in I} |\langle u_i, S^{-1}u\rangle|^2 + \sum_{i\in I} |\langle u_i, S^{-1}u\rangle - q_i|^2.$$

In particular, the frame coefficients $\{\langle u_i, S^{-1}u\rangle\}_{i\in I}$ of u with respect to the frame $\{u_i\}_{i\in I}$ represent the representation with minimal $\ell^2(I, \mathbb{H})$-norm."

Proof: We have $\sum_{i\in I} u_i \langle u_i, S^{-1}u\rangle = \sum_{i\in I} u_i q_i$, then, by multiplying both terms by $S^{-1}u$ on the left (in the sense of the inner product), we obtain:

$$\sum_{i\in I} |\langle u_i, S^{-1}u\rangle|^2 = \sum_{i\in I} \overline{\langle u_i, S^{-1}u\rangle} q_i.$$

On the other hand, we have:

$$\sum_{i\in I} |\langle u_i, S^{-1}u\rangle - q_i|^2 = \sum_{i\in I} |\langle u_i, S^{-1}u\rangle|^2 - 2\mathrm{Re}(\sum_{i\in I} \overline{\langle u_i, S^{-1}u\rangle} q_i) + \sum_{i\in I} |q_i|^2.$$

□

Corollary 3.4. *Let $\{u_i\}_{i\in I}$ be a frame of $\mathcal{H}$. Then for any $u \in \mathcal{H}$, we have:*

$$T^{\dagger}(u) = \left(\langle u_i, S^{-1}u\rangle\right)_{i\in I}.$$

That is:

$$T^{\dagger} = T^{*}S^{-1}.$$

Proof: T is a surjective right $\mathbb{H}$-linear bounded operator, then $R(T) = \mathcal{H}$. Let $u \in \mathcal{H}$ and consider the following equation $(E) : Tx = u$. By Theorem 2.15, The equation (E) has a unique solution with minimal norm which is exactely $T^\dagger u$. On the other hand, $q = \{q_i\}_{i\in I} \in \ell^2(I, \mathbb{H})$ is a solution to (E), if and only if, $\sum_{i\in I} u_i q_i = u$, then the unique solution with minimal norm of (E) is, by Theorem 3.3, $\{\langle u_i, S^{-1}u\rangle\}_{i\in I}$. Hence $T^\dagger u = \{\langle u_i, S^{-1}u\rangle\}_{i\in I} = T^* S^{-1}u$. □

We define the *optimal lower frame bound* of a frame as the supremum of all constants A satisfying the left-hand inequality in Definition 1.12, and the *optimal upper frame bound* as the infimum of all constants B satisfying the right-hand inequality in Definition 1.12. Note that both A_{opt} and B_{opt} are themselves valid frame bounds.

Therefore, for a given frame, the set of all lower frame bounds is the interval $]0, A_{\text{opt}}]$, and the set of all upper frame bounds is $[B_{\text{opt}}, +\infty[$.

According to Definition 1.12, the optimal frame bounds are given in terms of the associated frame operator S by:

$$A_{\text{opt}} := \inf_{\|u\|=1} \langle Su, u\rangle, \quad B_{\text{opt}} := \sup_{\|u\|=1} \langle Su, u\rangle.$$

The following theorem expresses the optimal frame bounds of a frame using its associated operators.

Theorem 3.5. *Let $\{u_i\}_{i\in I} \subset \mathcal{H}$ be a frame, T and S be, respectively, its pre-frame operator and its frame operator. Let $A_{opt} \leq B_{opt}$ be the optimal frame bounds of $\{u_i\}_{i\in I}$. Then:*

(1) $A_{opt} = \dfrac{1}{\|S^{-1}\|} = \dfrac{1}{\|T^\dagger\|^2}$.

(2) $B_{opt} = \|S\| = \|T\|^2$.

Proof: We have $\langle Su, u\rangle = \langle S^{\frac{1}{2}}u, S^{\frac{1}{2}}u\rangle = \|S^{\frac{1}{2}}u\|^2$, then $B_{opt} = \|S^{\frac{1}{2}}\|^2 = \|S\|$ and $A_{opt} = \dfrac{1}{\|S^{\frac{-1}{2}}\|^2} = \dfrac{1}{\|S^{-1}\|}$. On the other hand, we have $\|S\| = \|TT^*\| = \|T\|^2$ and since $T^\dagger = T^* S^{-1}$, then $\|T^\dagger\|^2 = \|T^{\dagger^*} T^\dagger\| = \|S^{-1}TT^*S^{-1}\| = \|S^{-1}SS^{-1}\| = \|S^{-1}\|$. □

Let $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$ with frame bounds $A \leq B$. If $R \in \mathbb{B}(\mathcal{H})$ and $u \in \mathcal{H}$, the frame coefficients of Ru are $\{\langle u_i, S^{-1}Ru\rangle\}_{i\in I}$. An interesting question arises: Can we determine the frame coefficients of Ru from those of u?

We consider the map defined as follows:

$$\begin{array}{rcl} \Lambda : \ell^2(I, \mathbb{H}) & \to & \ell^2(I, \mathbb{H}) \\ \{q_i\}_{i\in I} & \mapsto & \left\{\sum_{i\in I} \langle S^{-1}u_n, Ru_i\rangle q_i\right\}_{n\in I} . \end{array}$$

Proposition 3.6. *Λ is a well defined right $\mathbb{H}$-linear bounded operator. i.e., $\Lambda \in \mathbb{B}(\mathcal{H})$.*

Proof: Let $\{q_i\}_{i\in I} \in \ell^2(I, \mathbb{H})$. It is clear that, for all $n \in I$, $\sum_{i\in I} \langle S^{-1}u_n, Ru_i\rangle q_i \in \mathbb{H}$ since $\{\langle S^{-1}u_n, Ru_i\rangle\}_{i\in I}$, $\{q_i\}_{i\in I} \in \ell^2(I, \mathbb{H})$. Let $J \subset I$ be a finite subset of I. Then:

$$\begin{array}{rcl} \sum_{n\in J} |\sum_{i\in I} \langle S^{-1}u_n, Ru_i\rangle q_i|^2 & = & \sum_{n\in J} |\langle S^{-1}u_n, RT(\{q_i\}_{i\in I})\rangle|^2 = \sum_{n\in J} |\langle u_n, S^{-1}RT(\{q_i\}_{i\in I})\rangle|^2 \\ & \leq & B\|S^{-1}RT(\{q_i\}_{i\in I})\|^2 \leq B\|S^{-1}\|^2\|R\|^2\|T\|^2\|\{q_i\}_{i\in I}\| \\ & = & \left(\dfrac{B\|R\|}{A}\right)^2 \|\{q_i\}_{i\in I}\|^2 . \end{array}$$

Hence, Λ is well defined, clearly right $\mathbb{H}$-linear and bounded operator, moreover $\|\Lambda\| \leq \dfrac{B\|R\|}{A}$. □

The answer to the above question is given in the following proposition.

Proposition 3.7. *For all $u \in \mathcal{H}$, the frame coefficients of Ru are obtained from those of u via the right $\mathbb{H}$-linear bounded operator Λ.*

Proof: For $u \in \mathcal{H}$, we have:

$$\begin{aligned}\Lambda\left(\{\langle u_i, S^{-1}u\rangle\}_{i\in I}\right) &= \left(\sum_{i\in I}\langle S^{-1}u_n, Ru_i\rangle\langle u_i, S^{-1}u\rangle\right)_{n\in I} = \left(\left\langle S^{-1}u_n, R(\sum_{i\in I} u_i\langle u_i, S^{-1}u\rangle)\right\rangle\right)_{n\in I} \\ &= \left(\langle S^{-1}u_n, Ru\rangle\right)_{n\in I} = \left(\langle u_n, S^{-1}Ru\rangle\right)_{n\in I}.\end{aligned}$$

□

4. Frames and operators in quaternionic Hilbert spaces

In all this section $\mathcal{H}$ is a right quaternionic Hilbert space, $\mathbb{B}(\mathcal{H})$ is the set of all right $\mathbb{H}$-linear bounded operators on $\mathcal{H}$ and I is a countable set. In what follows, we characterize frames in a right quaternionic Hilbert space by the associated operators. The frames of the form $\{Lu_i\}_{i\in I}$, where $L \in \mathbb{B}(\mathcal{H})$ and $\{u_i\}_{i\in I}$ is a frame for $\mathcal{H}$, are studied.

We first give a characterization of frames by the pre-frame operator.

Theorem 4.1. *Let $\{u_i\}_{i\in I} \subset \mathcal{H}$ and T be its pre-frame operator. Then, the following statements are equivalent:*

(1) *$\{u_i\}_{i\in I}$ is a frame for $\mathcal{H}$.*

(2) *T is well defined, bounded and surjective.*

Proof: We have already seen that 1. implies 2. Conversely, since T is surjective, then, by Proposition 2.13, T^* (which is well defined and bounded since T is bounded) is bounded below. The fact that $\sum_{i\in I}|\langle u_i, u\rangle|^2 = \|T^*u\|^2$ completes the proof. □

In the following result, we characterize frames using the frame operator S under very weak conditions on the operator.

Theorem 4.2. *Let $\{u_i\}_{i\in I} \subset \mathcal{H}$ and S be its frame operator. Then, the following statements are equivalent:*

(1) *$\{u_i\}_{i\in I}$ is a frame for $\mathcal{H}$.*

(2) *S is well defined and surjective.*

Proof: It is well known that 1. implies 2.. Assume that:

$$\begin{aligned} S : \mathcal{H} &\to \mathcal{H} \\ u &\mapsto \sum_{i\in I} u_i\langle u_i, u\rangle, \end{aligned}$$

is well defined and surjective. Let $\{w_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$ such that

$$w_n \xrightarrow[n\to+\infty]{} w \in \mathcal{H} \quad \text{and} \quad Sw_n \xrightarrow[n\to+\infty]{} w' \in \mathcal{H}.$$

Since S is surjective, then there exists $v \in \mathcal{H}$ such that $w' = Sv$. Then:

$$\begin{aligned} Sw_n \xrightarrow[n\to+\infty]{} Sv \text{ and } w_n \xrightarrow[n\to+\infty]{} w &\implies \langle S(w_n - v), w_n - v\rangle \xrightarrow[n\to+\infty]{} 0 \text{ (Cauchy-Schwarz inequality)} \\ &\implies \sum_{i\in I}|\langle u_i, w_n - v\rangle|^2 \xrightarrow[n\to+\infty]{} 0 \text{ (by definition of } S) \\ &\implies \langle u_i, w_n - v\rangle \xrightarrow[n\to+\infty]{} 0 \ (\forall i \in I) \\ &\implies \langle u_i, w_n\rangle \xrightarrow[n\to+\infty]{} \langle u_i, v\rangle \\ &\implies \langle u_i, w\rangle = \langle u_i, v\rangle \ (\forall i \in I) \\ &\implies \sum_{i\in I} u_i\langle u_i, w\rangle = \sum_{i\in I} u_i\langle u_i, v\rangle \\ &\implies S(w) = S(v) = w'. \end{aligned}$$

Then, the graph of S is closed, hence, by the closed graph Theorem 2.5, S is bounded. Since S is positive and surjective, then it is invertible. Thus, $S^{\frac{1}{2}}$ is positive and invertible. Hence the existence of $0 < A \leq B$ such that for all $u \in \mathcal{H}$, $A\|u\|^2 \leq \|S^{\frac{1}{2}}u\|^2 \leq B\|u\|^2$. The fact that $\|S^{\frac{1}{2}}u\|^2 = \langle Su, u\rangle = \sum_{i\in I} |\langle u_i, u\rangle|^2$ completes the proof. □

Let L be a right $\mathbb{H}$-linear bounded operator and $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$. In what follows, we study the sequence $\{Lu_i\}_{i\in I}$.

In general, the image of a frame under a bounded operator (even if it is injective) is not a frame, as illustrated by the following example: Consider $\{v_i\}_{i\geqslant 1}$, a Hilbert basis of $\mathcal{H}$, and the following injective bounded linear operator:

$$\begin{array}{rcl} L : \mathcal{H} & \to & \mathcal{H} \\ x & \mapsto & \sum\limits_{i=1}^{+\infty} v_{i+1}\langle v_i, x\rangle. \end{array}$$

We have $\{L(v_i)\}_{i\geqslant 1} = \{v_i\}_{i\geqslant 2}$, which is not even complete.

The following proposition expresses the frame operator of $\{Lu_i\}_{i\in I}$ in terms of those associated with $\{u_i\}_{i\in I}$.

Proposition 4.3. *Let $L \in \mathbb{B}(\mathcal{H})$ and $\{u_i\}_{i\in I}$ be a Bessel sequence for $\mathcal{H}$. Then, $\{Lu_i\}_{i\in I}$ is a Bessel sequence for $\mathcal{H}$ and the associated operators are as follows:*

(1) $T_L = L^*T$.

(2) $\theta_L = \theta L$.

(3) $S_L = L^*SL$.

Here, T_L, θ_L and S_L are the pre-frame operator, the transform operator and the frame operator associated with $\{Lu_i\}_{i\in I}$, respectively. T, θ and S are the associated operators with $\{u_i\}_{i\in I}$.

Proof: Let B be a Bessel bound associated with $\{u_i\}_{i\in I}$. For all $u \in \mathcal{H}$, we have

$$\sum_{i\in I} |\langle Lu_i, u\rangle|^2 = \sum_{i\in I} |\langle u_i, L^*u\rangle|^2 \leq B\|L^*u\|^2 \leq B\|L\|^2\|u\|^2.$$

Hence, $\{Lu_i\}_{i\in I}$ is a Bessel sequence for $\mathcal{H}$. Let $\{q_i\}_{i\in I} \in \ell^2(I, \mathbb{H})$, we have

$$T_L(\{q_i\}_{i\in I}) = \sum_{i\in I} Lu_iq_i = L\left(\sum_{i\in I} u_iq_i\right) = LT(\{q_i\}_{i\in I}).$$

Then,

$$T_L = LT.$$

The rest follows immediately from (1). □

The next theorem presents a necessary and sufficient condition on $L \in \mathbb{B}(\mathcal{H})$ for it to transform one frame into another.

Theorem 4.4. *Let $L \in \mathbb{B}(\mathcal{H})$ and $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$. Then, $\{Lu_i\}_{i\in I}$ is a frame for $\mathcal{H}$ if and only if L is surjective.*

Proof: Let $A \leq B$ be frame bounds associated with $\{u_i\}_{i\in I}$.

Assume that $\{Lu_i\}_{i\in I}$ is a frame for $\mathcal{H}$, and let $v \in \mathcal{H}$. Then,

$$v = \sum_{i\in I} Lu_i \left\langle (LSL^*)^{-1}u_i, u\right\rangle = L\left(\sum_{i\in I} u_i \left\langle (LSL^*)^{-1}u_i, u\right\rangle\right).$$

Set

$$u := \sum_{i\in I} u_i \left\langle (LSL^*)^{-1}u_i, u\right\rangle \in \mathcal{H},$$

then $v = Lu$, which shows that L is surjective.

Conversely, by Proposition 2.13, L is surjective if and only if L^* is bounded below. Therefore, there exists a constant $M > 0$ such that for all $u \in \mathcal{H}$,

$$M\|u\| \leq \|L^*u\|.$$

Let $u \in \mathcal{H}$. Then:

$$\sum_{i\in I} |\langle Lu_i, u\rangle|^2 = \sum_{i\in I} |\langle u_i, L^*u\rangle|^2 .$$

It follows that:

$$A\|L^*u\|^2 \leq \sum_{i\in I} |\langle Lu_i, u\rangle|^2 \leq B\|L^*u\|^2.$$

Using the inequality $M\|u\| \leq \|L^*u\|$ and the bound $\|L^*u\| \leq \|L\|\|u\|$, we obtain:

$$AM^2\|u\|^2 \leq \sum_{i\in I} |\langle Lu_i, u\rangle|^2 \leq B\|L\|^2\|u\|^2.$$

Hence, $\{Lu_i\}_{i\in I}$ is a frame for $\mathcal{H}$ with frame bounds AM^2 and $B\|L\|^2$. □

Corollary 4.5. *Let $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$ with frame bounds $A \leq B$ and let S be its frame operator. Then:*

(1) *$\{S^{-1}u_i\}_{i\in I}$ is a frame for $\mathcal{H}$ with frame bounds $\frac{1}{B}$ and $\frac{1}{A}$. $\{S^{-1}u_i\}_{i\in I}$ is called the canonical dual frame for $\{u_i\}_{i\in I}$.*

(2) *$\{S^{\frac{-1}{2}}u_i\}_{i\in I}$ is a Parseval frame for $\mathcal{H}$.*

Proof: (1) By Proposition 4.3, Theorem 3.5, and Theorem 4.4, we easily deduce that the sequence $\{S^{-1}u_i\}_{i\in I}$ is a frame with optimal frame bounds given by

$$A'_{\text{opt}} = \frac{1}{B_{\text{opt}}} \quad \text{and} \quad B'_{\text{opt}} = \frac{1}{A_{\text{opt}}},$$

where A_{opt} and B_{opt} are the optimal frame bounds of the sequence $\{u_i\}_{i\in I}$.

We now extend this result to arbitrary frame bounds. Indeed, let $A \leq B$ be any frame bounds associated with $\{u_i\}_{i\in I}$. Then we have $A \leq A_{\text{opt}}$ and $B \geq B_{\text{opt}}$, which implies

$$\frac{1}{A} \geq \frac{1}{A_{\text{opt}}} = B'_{\text{opt}}, \quad \frac{1}{B} \leq \frac{1}{B_{\text{opt}}} = A'_{\text{opt}}.$$

Therefore, $\frac{1}{A}$ and $\frac{1}{B}$ are valid frame bounds for the sequence $\{S^{-1}u_i\}_{i\in I}$.

(2) Again, by Proposition 4.3, Theorem 3.5, and Theorem 4.4, we deduce that the sequence $\{S^{-\frac{1}{2}}u_i\}_{i\in I}$ is a frame for $\mathcal{H}$ with optimal frame bounds equal to 1. Hence, it is a Parseval frame for $\mathcal{H}$. □

The following result shows that unitary right $\mathbb{H}$-linear bounded operators transform frames to other frames with the same frame bounds.

Proposition 4.6. *Let $\{u_i\}_{i\in I}$ be a frame for $\mathcal{H}$, S be its frame operator and let $U \in \mathbb{B}(\mathcal{H})$ be a unitary right $\mathbb{H}$-linear bounded operator. Then $\{Uu_i\}_{i\in I}$ is a frame for $\mathcal{H}$ with the same frame bounds.*

Proof: By Theorem 4.4, $\{Uu_i\}_{i\in I}$ is a frame for $\mathcal{H}$. It sufficies to show that $\{u_i\}_{i\in I}$ and $\{Uu_i\}_{i\in I}$ have the same optimal frame bounds. i.e. in view of Theorem 3.5 and Proposition 4.3, we will show that $\|USU^*\| = \|S\|$ and $\|(USU^*)^{-1}\| = \|S^{-1}\|$. We have:

$$\begin{aligned} \|USU^*\| &= \sup_{\|u\|=1} \|USU^*(u)\| = \sup_{\|u\|=1} \|SU^*(u)\| = \|SU^*\| = \|US\| \\ &= \sup_{\|u\|=1} \|US(u)\| = \sup_{\|u\|=1} \|S(u)\| = \|S\|. \end{aligned}$$

And:

$$\begin{aligned}\|(USU^*)^{-1}\| = \|US^{-1}U^*\| &= \sup_{\|u\|=1}\|US^{-1}U^*(u)\| = \sup_{\|u\|=1}\|S^{-1}U^*(u)\| \\ &= \|S^{-1}U^*\| = \|US^{-1}\| = \sup_{\|u\|=1}\|US^{-1}(u)\| = \sup_{\|u\|=1}\|S^{-1}(u)\| = \|S^{-1}\|.\end{aligned}$$

□

In the following result, we study $\{Pu_i\}_{i\in I}$ where $\{u_i\}_{i\in I}$ is a frame for $\mathcal{H}$ and P is an orthogonal projection of $\mathcal{H}$.

Proposition 4.7. *Let $\{u_i\}_{i\in I}$ be a frame of $\mathcal{H}$ with frame bounds A and B, and let P be an orthogonal projection of $\mathcal{H}$. Then $\{Pu_i\}_{i\in I}$ is a frame of $P(\mathcal{H})$ with bounds A and B.*

Proof: Let $u \in P(\mathcal{H})$, then $Pu = u$ and $\langle Pu_i, u\rangle = \langle u_i, Pu\rangle = \langle u_i, u\rangle$. □

Corollary 4.8. *An orthogonal projection transforms a Parseval frame into a Parseval frame.*

The following theorem provides conditions under which one frame is the image of another under a right $\mathbb{H}$-linear bounded operator.

Theorem 4.9. *Let $\{u_i\}_{i\in I}$ and $\{v_i\}_{i\in I}$ be two frames of $\mathcal{H}$ with pre-frame operators T_1 and T_2, respectively.*

Define an operator L_0 on $\operatorname{span}\{u_i\}_{i\in I}$ *by*

$$L_0\left(\sum_{i\in I} u_i q_i\right) = \sum_{i\in I} v_i q_i$$

for any sequence $\{q_i\}_{i\in I} \subset \mathbb{H}$ that is zero except possibly for a finite number of elements.

Then, the following statements are equivalent:

(1) *L_0 is well defined as a right $\mathbb{H}$-linear bounded opeartor on* $\operatorname{span}\{u_i\}_{i\in I}$.

(2) $\ker(T_1) \subset \ker(T_2)$.

Proof: Denote by $\{e_i\}_{i\in I}$ the standard Hilbert basis of $\ell^2(I, \mathbb{H})$.

(1) $\Longrightarrow$ (2). Assume that L_0 is a right $\mathbb{H}$-linear bounded operator on $\operatorname{span}\{u_i\}_{i\in I}$, then it can be uniquely extended to a right $\mathbb{H}$-linear bounded operator on $\mathcal{H}$, denoted also by L_0, since $\operatorname{span}\{u_i\}_{i\in I}$ is dense in $\mathcal{H}$. Let $q := \{q_i\}_{i\in I} \subset \ell^2(I, \mathbb{H})$, we have:

$$\begin{aligned}T_1(q) = 0 &\Longrightarrow \sum_{i\in I} u_i\langle e_i, q\rangle = 0 \\ &\Longrightarrow L_0\left(\sum_{i\in I} u_i\langle e_i, q\rangle\right) = 0 \\ &\Longrightarrow \sum_{i\in I} L_0 u_i\langle e_i, q\rangle = 0 \\ &\Longrightarrow \sum_{i\in I} v_i\langle e_i, q\rangle = 0 \\ &\Longrightarrow T_2(q) = 0.\end{aligned}$$

Hence, $\operatorname{Ker}(T_1) \subset \operatorname{Ker}(T_2)$.

Conversely, assume that $\ker(T_1) \subset \ker(T_2)$. Let's show, first, that L_0 is well defined. Let $p := \{p_i\}_{i\in I},\ q := \{q_i\}_{i\in I} \subset \mathbb{H}$ be zero except possibly for a finite number of elements. We have:

$$\begin{aligned}\sum_{i\in I} u_i p_i = \sum_{i\in I} u_i q_i &\Longrightarrow T_1(p) = T_1(q) \\ &\Longrightarrow T_1(p-q) = 0 \\ &\Longrightarrow T_2(p-q) = 0 \\ &\Longrightarrow T_2(p) = T_2(q) \\ &\Longrightarrow \sum_{i\in I} v_i p_i = \sum_{i\in I} v_i q_i\end{aligned}$$

Hence, L_0 is well defined from $\text{span}\{u_i\}_{i\in I}$ to $\text{span}\{v_i\}_{i\in I}$. It is clear that L_0 is right $\mathbb{H}$-linear. Let's show now that L_0 is bounded.

Let $T = T_1$ or T_2, it is clear that $T_{|\ker(T)^\perp} : \ker(T)^\perp \to \mathcal{H}$ is invertible since $T : \ell^2(\mathbb{H}) \to \mathcal{H}$ is surjective. Then, by the open mapping theorem 2.4, $(T_{|\ker(T)^\perp})^{-1}$ is also bounded, and then for all $q \in \ker(T)^\perp$, we have:

$$\alpha\|q\| \leq \|T_{|\ker(T)^\perp}(q)\| \leq \beta\|q\|,$$

where $\alpha = \dfrac{1}{\|(T_{|\ker(T)^\perp})^{-1}\|}$ and $\beta = \|T_{|\ker(T)^\perp}\|$.

Let now $q = \{q_i\}_{i\in I} \subset \mathbb{H}$ be zero except possibly for a finite number of elements and denote $q_1 := P_{\ker(T_1)^\perp} q$ and $q_2 := P_{\ker(T_2)^\perp} q$, where P_F is the orthogonal projection onto $F \subset \ell^2(\mathbb{H})$ closed. We have:

$$\begin{aligned}
\left\| L_0\left(\sum_{i\in I} u_i q_i\right)\right\| &= \left\|\sum_{i\in I} v_i q_i\right\| = \|T_{2|\ker(T_2)^\perp}(q_2)\| \leq \beta_2\|q_2\| \\
&\leq \beta_2\|q_1\| \quad \text{since } \ker(T_1) \subset \ker(T_2) \\
&\leq \frac{\beta_2}{\alpha_1}\|T_{1|\ker(T_1)^\perp}(q_1)\| = \frac{\beta_2}{\alpha_1}\|T_1(q)\| = \frac{\beta_2}{\alpha_1}\left\|\sum_{i\in I} u_i q_i\right\|
\end{aligned}$$

Then $L_0 : \text{span}\{u_i\}_{i\in I} \to \text{span}\{v_i\}_{i\in I}$ is bounded.

□

Definition 4.10. Let $\{u_i\}_{i\in I}$ and $\{v_i\}_{i\in I}$ be two sequences of vectors in $\mathcal{H}$. We say that $\{u_i\}_{i\in I}$ and $\{v_i\}_{i\in I}$ are equivalent, and we write $\{u_i\}_{i\in I} \approx \{v_i\}_{i\in I}$, if there exists a right $\mathbb{H}$-linear bounded invertible operator $L : \mathcal{H} \to \mathcal{H}$ such that for all $i \in I$, $L(u_i) = v_i$.

Remark 4.11. In view of the open map theorem 2.4, we have that:

$$\{u_i\}_{i\in I} \approx \{v_i\}_{i\in I} \Longleftrightarrow \{v_i\}_{i\in I} \approx \{u_i\}_{i\in I}.$$

Example 4.12. Each frame for $\mathcal{H}$ is equivalent to a Parseval frame for $\mathcal{H}$. Indeed, we have already seen, in Corollary 4.5, that if $\{u_i\}_{i\in I}$ is a frame for $\mathcal{H}$ with the frame operator S, then $\{S^{\frac{-1}{2}} u_i\}_{i\in I}$ is a Parseval frame for $\mathcal{H}$.

The following theorem provides sufficient and necessary conditions under which two frames are equivalent.

Theorem 4.13. *Let $(u_i)_{i\in I}$ and $(v_i)_{i\in I}$ be two frames of $\mathcal{H}$ with pre-frame operators T_1 and T_2, respectively. Then the following statements are equivalent.*

(1) $\{u_i\}_{i\in I} \approx \{v_i\}_{i\in I}$.

(2) $ker(T_1) = ker(T_2)$.

Proof: Assume that $\{u_i\}_{i\in I} \approx \{v_i\}_{i\in I}$, then there exists a right $\mathbb{H}$-linear, bounded and invertible operator L such that $Lu_i = v_i$ for all $i \in I$. Then, by Theorem 4.9, $ker(T_1) \subset ker(T_2)$. By the open map theorem 2.4, L^{-1} is also a right $\mathbb{H}$-linear bounded operator and since $v_i = L^{-1}u_i$, then by Theorem 4.9, $ker(T_2) \subset ker(T_1)$. Hence $ker(T_1) = ker(T_2)$. Conversely, assume that $ker(T_1) = ker(T_2)$, then by Theorem 4.9, $L_1 : u_i \to v_i$ and $L_2 : v_i \to u_i$ are two right $\mathbb{H}$-linear bounded operators. Moreover, $L_1L_2 = L_2L_1 = I$, where I is the identity operator of $\mathbb{B}(\mathcal{H})$. Hence $\{u_i\}_{i\in I}$ and $\{v_i\}_{i\in I}$ are equivalent. □

We have seen that each frame for $\mathcal{H}$ defines a positive and invertible right $\mathbb{H}$-linear bounded operator which is its associated frame operator. The next theorem investigates the reciproque, i.e., Given a positive invertible operator $S_0 \in \mathbb{B}(\mathcal{H})$, is there a frame for $\mathcal{H}$ for which the frame operator is S_0.

Theorem 4.14. *Let $S_0 \in \mathbb{B}(\mathcal{H})$ a positive and invertible operator on $\mathcal{H}$. Then, there exists a frame for $\mathcal{H}$ for which the frame operator is S_0.*

Proof: Let $\{v_i\}_{i\in I}$ be a Hilbert basis for $\mathcal{H}$ and define for all $i \in I$, $u_i := S_0^{\frac{1}{2}} v_i$. By Theorem 4.4, $\{u_i\}_{i\in I}$ is a frame for $\mathcal{H}$ and, by Proposition 4.3, its frame opeartor is $S_0^{\frac{1}{2}} \operatorname{Id} S_0^{\frac{1}{2}} = S_0$, where Id is the identity operator of $\mathbb{B}(\mathcal{H})$. □

The next theorem shows that the correspondence between Bessel sequences and $\mathbb{H}$-linear bounded operators is bijective.

Theorem 4.15. *Let $\mathcal{B}$ denote the set of all Bessel sequences in $\mathcal{H}$ indexed by a countable set I. Then the map:*

$$\begin{array}{rcl} \Theta : \mathcal{B} & \to & \mathbb{B}(\mathcal{H}, \ell^2(\mathbb{H})) \\ \{u_i\}_{i\in I} & \mapsto & \theta_u, \end{array}$$

where θ_u is the transform operator of $\{u_i\}_{i\in I}$, is bijective. Moreover,

$$\begin{array}{rcl} \Theta^{-1} : \mathbb{B}(\mathcal{H}, \ell^2(\mathbb{H})) & \to & \mathcal{B} \\ L & \mapsto & \{L^*(v_i)\}_{i\in I}, \end{array}$$

where $\{v_i\}_{i\in I}$ is the standard Hilbert basis for $\ell^2(\mathbb{H})$.

Proof: Let $u = \{u_i\}_{i\in I}$, $w = \{w_i\}_{i\in I}$ be two Bessel sequences for $\mathcal{H}$. We have:

$$\begin{array}{rcl} \theta_u = \theta_w & \Longrightarrow & \forall x \in \mathcal{H},\ \forall i \in I,\ \langle u_i, x\rangle = \langle w_i, x\rangle \\ & \Longrightarrow & \forall i \in I,\ u_i = w_i \\ & \Longrightarrow & u = w. \end{array}$$

Then, Θ is injective. Let $L \in \mathbb{B}(\mathcal{H}, \ell^2(\mathbb{H}))$, and set $f = \{L^* v_i\}_{i\in I}$. We have for all $u \in \mathcal{H}$, $\theta_f u = \{\langle L^* v_i, u\rangle\}_{i\in I} = \{\langle v_i, Lu\rangle\}_{i\in I} = Lu$ since $\{v_i\}_{i\in I}$ is the standard Hilbert basis for $\ell^2(\mathbb{H})$. Hence, Θ is surjective, thus is bijective. Moreover:

$$\begin{array}{rcl} \Theta^{-1} : \mathbb{B}(\mathcal{H}, \ell^2(\mathbb{H})) & \to & \mathcal{B} \\ L & \mapsto & \{L^*(v_i)\}_{i\in I}, \end{array}$$

where $\{v_i\}_{i\in I}$ is the standard Hilbert basis for $\ell^2(\mathbb{H})$. □

Remark 4.16. The set $\mathbb{B}(\mathcal{H}, \ell^2(\mathbb{H}))$ is a $\mathbb{R}$-vector space and the well known norm of right $\mathbb{H}$-linear bounded operators is a $\mathbb{R}$-norm on $\mathbb{B}(\mathcal{H}, \ell^2(\mathbb{H}))$. We set for for all $u := \{u_i\}_{i\in I} \in \mathcal{B}$, $\|u\| := \|\theta_u\|$, then Θ is a $\mathbb{R}$-linear, bounded invertible operator. Moreover, Θ is isometric.

Acknowledgments

It is my great pleasure to thank the anonymous referees for their careful reading of the paper and for several valuable suggestions.

References

[1] Ghiloni R., Moretti V., Perotti A., Continuous slice functional calculus in quaternionic Hilbert spaces. Reviews in Mathematical Physics, (2013) Vol. 25, P. 1350006.

[2] Sharma S. K., Goel S., Frames in quaternionic Hilbert spaces. Journal of Mathematical Physics, Analysis, Geometry, (2019) Vol. 15, P. 395–411.

Najib Khachiaa,
University Ibn Tofail, Kenitra, Morocco
e-mail: khachiaa.najib@uit.ac.ma

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 136–144
DOI: 10.29229/uzmj.2026-2-16

Optimization of the Euler-Maclaurin type quadrature formula in the Hilbert space of periodic functions

Khayriev U., Nutfullaeva A., Hamroeva Z.

Abstract. In this paper, we address the construction of an optimal quadrature formula in the sense of Sard within a Hilbert space composed of periodic, complex-valued functions, specifically for the purpose of numerically approximating Fourier integrals. The quadrature formula is expressed as a linear combination of the function's values taken at nodes of a uniform grid. An upper estimate of the quadrature error is obtained through the norm of the associated error functional, which is evaluated using the Cauchy-Schwarz inequality. To determine this norm explicitly, the method relies on the concept of an extremal function. The extremal function corresponding to the error functional is identified by applying the Riesz representation theorem, which guarantees that the norm of this function coincides with the norm of the error functional in the dual (conjugate) space. Ultimately, the resulting expression for the norm of the error functional takes the form of a multivariate quadratic function dependent on the coefficients of the quadrature formula. In addition, this work determines the optimal coefficients of a derived quadrature formula. Using these optimal coefficients, the exponential-weighted integrals of several functions have been approximately evaluated.

Keywords: Hilbert space, extremal function, error functional, optimal quadrature formula, strongly oscillatory integrals, Fourier transform
MSC (2020): 41A05, 41A15, 65D30, 65D32

1. Introduction and statement of the problem

1.1. **Introduction.** The numerical computation of strongly oscillating integrals is a critical problem in numerical analysis, as such integrals are extensively applied in science and technology. The following types of Fourier integrals also serve as examples of strongly oscillating integrals for sufficiently large values of $\omega \in \mathbb{R}$

$$I(\omega, \varphi) = \int_0^1 e^{2\pi i \omega x} \varphi(x) dx,$$

where φ is non-oscillating functions.

The numerical approximation of these integrals can be particularly challenging when the oscillations are highly localized and rapidly alternating. However, several techniques exist to achieve accurate numerical approximations of such integrals. One such method is the stationary phase method, which involves identifying the stationary points of the integrand and approximating the integral based on the behaviour of the integrand near these points. Another approach is to apply a quadrature method specifically designed for oscillatory integrals, such as the Filon method, Clenshaw-Curtis method, modified Clenshaw-Curtis method, Levin-type methods, Gauss-Laguerre quadrature, and generalized quadrature methods (for comprehensive details, see [1], for example, [2, 3] and the references therein).

In recent years, Kh.M. Shadimetov [4, 5, 6, 7, 8], G.V. Milovanović [2, 1], and A.R. Hayotov [9, 10, 11, 12, 13], N.D. Boltaev [14], I.O. Jalolov [15, 16] and S.S. Babaev [17, 18, 19] have conducted scientific research focused on constructing optimal quadrature formulas for calculating Fourier coefficients and integrals in various spaces, including $L_2^{(m)}$ and $W_2^{(m,m-1)}$. The outcomes of constructing optimal quadrature formulas for the numerical evaluation of Fourier coefficients in $\widetilde{W}_2^{(2,1)}$ space of periodic functions, as well as applying these formulas to reconstruct computed tomography images, were obtained in the studies of A.R. Hayotov [20]. It can also be added that optimal quadrature formulas for the numerical calculation of singular integrals in the spaces $L_2^{(m)}$ and $W_2^{(m,m-1)}$ are constructed in the following works [21, 22, 23].

In this paper, the problem of constructing an optimal quadrature formula of Euler-Maclaurin type in the Hilbert space $\widetilde{W}_2^{(2,1)}$ of periodic functions is studied.

1.2. **Statement of the problem.** This study focuses on the derivation of optimal quadrature formulas under the assumption that the integrand lies within the Hilbert space $W_2^{(2,1)}$. For a detailed description of this function space, the reader is referred to the definition provided, for instance, in the references [14, 24, 25, 26].

The space $W_2^{(2,1)}$ is a Hilbert space consisting of complex-valued functions, and it is defined as follows

$$W_2{}^{(2,1)} = \{\varphi : [0,1] \to \mathbb{C} | \varphi' \text{ is abs. continuous and } \varphi'' \in L_2(0,1)\}.$$

The space is the Hilbert space with the inner product

$$\langle \varphi, \psi \rangle_{W_2^{(2,1)}} = \int_0^1 (\varphi''(x) + \varphi'(x))(\overline{\psi}''(x) + \overline{\psi}'(x))dx, \tag{1.1}$$

and the corresponding norm is

$$\|\varphi\|_{W_2^{(2,1)}} = \left(\langle \varphi, \varphi \rangle_{W_2^{(2,1)}}\right)^{1/2}.$$

Let $\widetilde{W}_2^{(2,1)}(0,1]$ denote the subspace of $W_2^{(2,1)}(0,1)$ that consists of 1-periodic complex-valued functions. Each function belonging to $\widetilde{W}_2^{(2,1)}(0,1]$ satisfies the condition of 1-periodicity

$$\varphi(x+\beta) = \varphi(x) \text{ for } x \in \mathbb{R} \text{ and } \beta \in \mathbb{Z}.$$

We consider the quadrature formula of the following form

$$\int_0^1 e^{2\pi i w x} \varphi(x) dx \cong \sum_{k=1}^{N} C_k \varphi(hk) + \sum_{k=1}^{N} A_k \varphi'(hk), \tag{1.2}$$

where $\omega \in \mathbb{Z} \setminus \{0\}$, $\varphi \in \widetilde{W}_2{}^{(2,1)}$, C_k coefficients are given in the work [5]

$$C_k = \frac{2}{4\pi^2\omega^2 + 1} \cdot \frac{e^{2h} - 2e^h \cos(2\pi\omega h) + 1}{e^{2h} - 1} \cdot e^{2\pi i \omega h k} \quad \text{for} \quad k = 1, 2, \ldots, N,$$

the coefficients A_k are unknowns and need to be determined, while h represents the step size of the equal mesh. The difference between the integral and the quadrature sum is called *the error* of the quadrature formula (1.2)

$$(\ell, \varphi) = \int_0^1 e^{2\pi i w x} \varphi(x) dx - \sum_{k=1}^{N} C_k \varphi(hk) - \sum_{k=1}^{N} A_k \varphi'(hk), \tag{1.3}$$

and the corresponding *error functional* is

$$\ell(x) = \left(e^{2\pi i w x} \varepsilon_{(0,1]}(x) - \sum_{k=1}^{N} C_k \delta(x - hk) + \sum_{k=1}^{N} A_k \delta'(x - hk) \right) * \phi_0(x), \tag{1.4}$$

where $\varepsilon_{(0,1]}(x)$ is the characteristic function of the interval $(0,1]$, δ is the Dirac delta-function and $\phi_0(x) = \sum\limits_{\beta=-\infty}^{\infty} \delta(x-\beta)$. The error functional ℓ is called the periodic error functional of the quadrature formula (1.2), and it belongs to the conjugate space $\widetilde{W}_2^{(2,1)*}(0,1]$.

Since the error functional ℓ is defined on the space $\widetilde{W}_2^{(2,1)}(0,1]$, the following equality is valid as in the work [27]

$$(\ell, 1) = 0. \tag{1.5}$$

The error (1.3) of the quadrature formula (1.3) is a linear functional in $\widetilde{W}_2^{(2,1)*}(0,1]$. The absolute value of the error (1.3) is estimated by the Cauchy-Schwarz inequality as follows

$$|(\ell, \varphi)| \le \|\ell\|_{\widetilde{W}_2{}^{(2,1)*}} \cdot \|\varphi\|_{\widetilde{W}_2{}^{(2,1)}},$$

where

$$\|\ell\|_{\widetilde{W_2}^{(2,1)*}} = \sup_{\varphi,\ \|\varphi\|_{\widetilde{W_2}^{(2,1)}}\neq 0} \frac{|(\ell,\varphi)|}{\|\varphi\|}$$

is the norm of the error functional (1.4).

Hence, to get the least upper bound of the error for the quadrature formula (1.2), we solve the following problems.

Problem 1.1. *Find the norm of the error functional* (1.4) *of the quadrature formula* (1.2) *in the space* $\widetilde{W}_2^{(2,1)*}(0,1]$.

It is clear that the norm of the error functional ℓ depends on the coefficients A_k. Minimizing $\|\ell\|$ by adjusting the coefficients A_k is a simple linear problem.

Problem 1.2. *Find the coefficients* $\overset{\circ}{A}_k$ *that minimize the value of the norm of the error functional* (1.4).

The coefficients that give $\|\ell\|_{\widetilde{W_2}^{(2,1)*}}$ the minimum value are called *optimal coefficients* and are denoted by $\overset{\circ}{A}_k$. The quadrature formula (1.2) with these coefficients is *the optimal quadrature formula in the sense of Sard.* Similar problem as Problem 1.1 was first proposed by S.L. Sobolev [27], later the problem was solved in the space $W_2^{(m,m-1)}$ of functions in works [9] and in the space $\widetilde{L_2^{(m)}}$ of periodic functions in works [28] and in works [28, 4] for optimal quadrature formulas of Euler-Maclaurin type.

Furthermore, to solve Problems 1.1 and 1.2 in the next sections, we will do the following:

- first, we find the extremal function of the quadrature formula (1.2) and use it to find the analytic representation of the norm for the error functional (1.4) in section 2;
- in section 3, the process of solving the system of linear equations for the coefficients of the optimal quadrature formula is presented;
- and in the last section, we find the sharp upper bound of the error of the constructed optimal quadrature formula.

2. Finding the norm of the error functional

In this section, we will deal with Problem 1.1. To do this, we use the concept of an extremal function.

2.1. **The extremal function.** To calculate the norm $\|\ell\|_{\widetilde{W_2}^{(2,1)*}}$, we use *the extremal function* ψ_ℓ for the error functional ℓ (see [27]) that satisfies the following equality:

$$(\ell,\psi_\ell) = \|\ell\|_{\widetilde{W_2}^{(2,1)*}} \cdot \|\psi_\ell\|_{\widetilde{W_2}^{(2,1)}}. \tag{2.1}$$

Since $\widetilde{W_2}^{(2,1)}$ is the Hilbert space, using the Riesz representation theorem, we obtain

$$(\ell,\varphi) = \langle\varphi,\psi_\ell\rangle_{\widetilde{W_2}^{(2,1)}}, \tag{2.2}$$

where $\langle\psi_\ell,\varphi\rangle_{\widetilde{W_2}^{(2,1)}}$ is the inner product of the functions ψ_ℓ and φ which is defined by expression (1.1). In addition, the equality $\|\ell\|_{\widetilde{W_2}^{(2,1)*}} = \|\psi_\ell\|_{\widetilde{W_2}^{(2,1)}}$ is satisfied. Thus, taking into account (2.1), we obtain

$$(\ell,\psi_\ell) = \|\ell\|^2_{\widetilde{W_2}^{(2,1)*}}. \tag{2.3}$$

Then from (2.2), integrating by parts, we obtain

$$\int_0^1 \ell(x)\varphi(x)dx = \int_0^1 \left(\overline{\psi_\ell}^{IV}(x) - \overline{\psi_\ell}''(x)\right)\varphi(x)dx.$$

From the last equation, we have

$$\overline{\psi_\ell}^{IV}(x) - \overline{\psi_\ell}''(x) = \ell(x). \tag{2.4}$$

For the solution of the last equation the following holds.

Theorem 2.1. *The solution of differential equation* (2.4) *is identified as the extremal function* ψ_ℓ *associated with the error functional* ℓ*, and can be expressed in the following form:*

$$\psi_\ell(x) = d_0 + e^{-2\pi i\omega x}\cdot\kappa(\omega) - \sum_{k=1}^{N}\overline{C_k}\sum_{\beta\neq 0}\kappa(\beta)e^{2\pi i\beta(x-hk)} - \sum_{k=1}^{N}\overline{A_k}\sum_{\beta\neq 0}\kappa(\beta)\cdot(2\pi i\beta)e^{2\pi i\beta(x-hk)}, \tag{2.5}$$

where d_0 *is a complex number,* $\overline{C_k}$ *and* $\overline{A_k}$ *are the complex conjugates of* C_k, A_k*, respectively and*

$$\kappa(\omega) = \frac{1}{(2\pi\omega)^4 + (2\pi\omega)^2}. \tag{2.6}$$

Proof: We find the periodic solution of equation (2.4) using the following properties of the Fourier transform (see, for instance [14])

$$F[\varphi] = \int_{-\infty}^{\infty}\varphi(x)e^{2\pi ipx}dx, F^{-1}[\varphi] = \int_{-\infty}^{\infty}\varphi(p)e^{-2\pi ipx}dp,$$

$$F[\varphi^{(\alpha)}] = (-2\pi ip)^\alpha F[\varphi],\ (\alpha\in\mathbb{N}), F^{-1}[F[\varphi(x)]] = \varphi(x).$$

We apply the Fourier transform to both sides of equation (2.4)

$$F[\overline{\psi_\ell}^{IV} - \overline{\psi_\ell}''] = F[\ell].$$

Since the Fourier transform is a linear operator, we have

$$\left((2\pi ip)^4 - (2\pi ip)^2\right)F[\overline{\psi_\ell}] = F\left[\ell\right]. \tag{2.7}$$

To find the Fourier transform of the error functional ℓ, we simplify it

$$\ell(x) = \left(e^{2\pi iwx}\varepsilon_{(0,1]}(x) - \sum_{k=1}^{N}C_k\delta(x-hk) + \sum_{k=1}^{N}A_k\delta'(x-hk)\right) * \sum_{\beta=-\infty}^{\infty}\delta(x-\beta).$$

Taking into account the convolution of two continuous functions f and g, which is

$$f(x) * g(x) = \int_{-\infty}^{\infty} f(y)\cdot g(x-y)dy$$

and properties of the Dirac delta-function, we get the following

$$\ell(x) = e^{2\pi iwx} - \sum_{k=1}^{N}C_k\sum_{\beta=-\infty}^{\infty}\delta(x-hk-\beta) + \sum_{k=1}^{N}A_k\sum_{\beta=-\infty}^{\infty}\delta'(x-hk-\beta). \tag{2.8}$$

Using (2.8) we rewrite eqution (2.7)

$$\left((2\pi ip)^4 - (2\pi ip)^2\right)F[\overline{\psi_\ell}] = F\left[e^{2\pi iwx} - \sum_{k=1}^{N}C_k\sum_{\beta=-\infty}^{\infty}\delta(x-hk-\beta) + \sum_{k=1}^{N}A_k\sum_{\beta=-\infty}^{\infty}\delta'(x-hk-\beta)\right],$$

or

$$\begin{aligned}\left((2\pi ip)^4 - (2\pi ip)^2\right)F[\bar{\psi}_\ell] &= F\left[e^{2\pi i\omega x}\right] - \sum_{k=1}^{N}C_k\sum_{\beta=-\infty}^{\infty}F\left[\delta(x-hk-\beta)\right] \\ &+ \sum_{k=1}^{N}A_k\sum_{\beta=-\infty}^{\infty}F\left[\delta'(x-hk-\beta)\right].\end{aligned} \tag{2.9}$$

Now, using the following equalities

$$F\left[e^{2\pi i\omega x}\right]=\delta(p+\omega),\quad \sum_{\beta=-\infty}^{\infty}F\left[\delta'(x-hk-\beta)\right]=(-2\pi ip)e^{2\pi iphk}\sum_{\beta=-\infty}^{\infty}\delta(p-\beta),$$

$$\sum_{\beta=-\infty}^{\infty}e^{2\pi ip\beta}=\sum_{\beta=-\infty}^{\infty}\delta(p-\beta),$$

$$\sum_{\beta=-\infty}^{\infty}F\left[\delta(x-hk-\beta)\right]=\sum_{\beta=-\infty}^{\infty}e^{2\pi ip(hk+\beta)}=e^{2\pi iphk}\sum_{\beta=-\infty}^{\infty}e^{2\pi ip\beta}=e^{2\pi iphk}\sum_{\beta=-\infty}^{\infty}\delta(p-\beta)$$

we can rewrite equation (2.9) as follows

$$\left((2\pi ip)^4-(2\pi ip)^2\right)F[\bar{\psi}_\ell]=\delta(p+\omega)-\sum_{k=1}^{N}C_k\sum_{\beta=-\infty}^{\infty}e^{2\pi iphk}\delta(p-\beta)\ -\sum_{k=1}^{N}A_k\sum_{\beta=-\infty}^{\infty}(2\pi ip)e^{2\pi iphk}\delta(p-\beta).$$

We consider that the coefficient on the left-hand side of the last equation is not equal to zero. Consequently, we can divide both sides of the last equation by $\kappa(p)$, which is defined by (2.6). This division is not uniquely defined. From the last equation, the function $F[\bar{\psi}_\ell]$ is defined up to the term of the form $\delta(p)$. Taking into account the above-mentioned and the properties of the delta-function, we get

$$F[\bar{\psi}_\ell]=\delta(p+\omega)\kappa(p)-\sum_{k=1}^{N}C_k\sum_{\beta\neq0}e^{2\pi iphk}\delta(p-\beta)\kappa(p)\ -\sum_{k=1}^{N}A_k\sum_{\beta\neq0}(2\pi ip)e^{2\pi iphk}\kappa(p)+d_0\delta(p),$$

where $\kappa(p)$ is defined by formula (2.6) and d_0 is a constant.

Using the property $f(x)\delta(x-a)=f(a)\delta(x-a)$ of delta-function, we have the following

$$F[\bar{\psi}_\ell]=\delta(p+\omega)\kappa(\omega)-\sum_{k=1}^{N}C_k\sum_{\beta\neq0}e^{2\pi i\beta hk}\kappa(\beta)\delta(p-\beta)\ -\sum_{k=1}^{N}A_k\sum_{\beta\neq0}(2\pi i\beta)e^{2\pi i\beta hk}\kappa(\beta)\delta(p-\beta)+d_0\delta(p).$$

Then, applying the inverse Fourier transform to both sides of the last equation, we have

$$\bar{\psi}_\ell=e^{2\pi i\omega x}\kappa(\omega)-\sum_{k=1}^{N}C_k\sum_{\beta\neq0}e^{2\pi i\beta hk}\kappa(\beta)e^{-2\pi i\beta x}-\sum_{k=1}^{N}A_k\sum_{\beta\neq0}(2\pi i\beta)e^{2\pi i\beta hk}\kappa(\beta)e^{-2\pi i\beta x}+d_0.$$

Since $\overline{\bar{\psi}}_\ell$ is equal to ψ_ℓ, we obtain (2.5), that is, Theorem 2.1 is proved. □

2.2. **The norm of the error functional.** The following result corresponds to the solution to Problem 1.1.

Theorem 2.2. *Within the functional space* $\widetilde{W_2}^{(2,1)*}(0,1]$, *the squared norm* $\|\ell\|_{\widetilde{W_2}^{(2,1)*}}$ *corresponding to* $\omega\in\mathrm{Z}\backslash\{0\}$ *is represented by the following expression*

$$\begin{aligned}\|\ell\|^2_{\widetilde{W_2}^{(2,1)*}} &= \|\ell_2\|^2_{\widetilde{W_2}^{(2,1)*}}+2\pi i\omega\cdot\kappa(\omega)\left[\sum_{k=1}^{N}A_ke^{-2\pi i\omega hk}+\sum_{k'=1}^{N}\overline{A_{k'}}e^{2\pi i\omega hk'}\right]\\ &+ \sum_{k=1}^{N}\sum_{k'=1}^{N}\sum_{\beta\neq0}2\pi i\beta\cdot\kappa(\beta)e^{2\pi i\beta h(k-k')}\left[A_k\overline{C_{k'}}+C_k\overline{A_{k'}}\right]\\ &+ \sum_{k=1}^{N}\sum_{k'=1}^{N}A_k\overline{A_{k'}}\sum_{\beta\neq0}(2\pi i\beta)^2\cdot\kappa(\beta)e^{2\pi i\beta h(k-k')},\end{aligned}$$

where

$$\|\ell_2\|^2_{\widetilde{W_2}^{(2,1)*}}=\kappa(\omega)-\kappa(\omega)\left[\sum_{k=1}^{N}C_ke^{-2\pi i\omega hk}+\sum_{k'=1}^{N}\overline{C_{k'}}e^{2\pi i\omega hk'}\right]+\sum_{k=1}^{N}\sum_{k'=1}^{N}C_k\overline{C_{k'}}\sum_{\beta\neq0}\kappa(\beta)e^{2\pi i\beta h(k-k')}$$

and $\kappa(\cdot)$ *is defined by* (2.6).

Proof: To prove Theorem 2.2, we calculate the norm $\|\ell\|_{\widetilde{W}_2^{(2,1)*}}$ and use equalities (2.3), (2.8) and (2.5), respectively. As a result we have the following

$$\begin{aligned}\|\ell\|^2_{\widetilde{W}_2^{(2,1)*}} = (\ell,\psi_\ell) &= \int_0^1 \ell(x)\Bigg(d_0 + e^{-2\pi i\omega x}\cdot\kappa(\omega) - \sum_{k=1}^{N}\overline{C_k}\sum_{\beta\neq 0}\kappa(\beta)e^{2\pi i\beta(x-hk)}\\ &- \sum_{k=1}^{N}\overline{A_k}\sum_{\beta\neq 0}\kappa(\beta)\cdot(2\pi i\beta)e^{2\pi i\beta(x-hk)}\Bigg)dx.\end{aligned}$$

As such, given the algorithm to calculate the norm $\|\ell_0\|_{\widetilde{W}_2^{(2,1)*}}$ in the work [5].

By taking condition (1.5) into account and simplifying the preceding expression, we arrive at the result stated in the previous theorem. Thus, Theorem 2.2 is completely proved. □

3. THE OPTIMAL COEFFICIENTS OF THE QUADRATURE FORMULA

In this section, we determine the optimal coefficients of the quadrature formula (1.2) that minimize the norm of the error functional ℓ. To find the minimum of the norm $\|\ell\|_{\widetilde{W}_2^{(2,1)*}}$with respect to the coefficients A_k, we employ the method of Lagrange multipliers for constrained optimization. Accordingly, we consider the following function:

$$L(A_k,\overline{A_{k'}}) = \|\ell\|^2 \text{ for } k=1,2,\dots,N \quad \text{and} \quad k'=1,2,\dots,N.$$

By setting the partial derivatives of the function $L(A_k,\overline{A_{k'}})$ with respect to all variables A_k and $\overline{A_{k'}}$ equal to zero, we obtain the following system of equations

$$\begin{aligned}\frac{\partial L}{\partial A_k} &= 2\pi i\omega\cdot\kappa(\omega)e^{-2\pi i\omega hk} + \sum_{k'=1}^{N}\overline{C_{k'}}\sum_{\beta\neq 0}2\pi i\beta\kappa(\beta)e^{-2\pi i\beta h(k-k')}\\ &- \sum_{k'=1}^{N}\overline{A_{k'}}\sum_{\beta\neq 0}(2\pi\beta)^2\kappa(\beta)e^{-2\pi i\beta h(k-k')} = 0,\ \ k=1,2,\dots,N, \qquad (3.1)\\ \frac{\partial L}{\partial\overline{A_{k'}}} &= 2\pi i\omega\cdot\kappa(\omega)e^{2\pi i\omega hk'} + \sum_{k=1}^{N}C_k\sum_{\beta\neq 0}2\pi i\beta\cdot\kappa(\beta)e^{2\pi i\beta h(k-k')}\\ &- \sum_{k'=1}^{N}A_k\sum_{\beta\neq 0}(2\pi\beta)^2\kappa(\beta)e^{2\pi i\beta h(k-k')} = 0,\ \ k=1,2,\dots,N, \qquad (3.2)\end{aligned}$$

where $\kappa(\cdot)$ is defined by (2.6).

A solution of this system which we denote by $\mathring{A}_k$, is a minimum point for the function $L(A_k,\overline{A_{k'}})$. The system of equations (3.1) is equivalent to system (3.2). Therefore, it is sufficient to solve the system (3.1).
The following theorem is valid for the solution of the system (3.1).

Theorem 3.1. *Among all quadrature formulas of the form* (1.2)*, where* $\omega\in\mathbb{Z}\backslash\{0\}$ *and* $\omega h\notin\mathbb{Z}$*, within the space* $\widetilde{W}_2^{(2,1)}(0,1]$*of complex-valued periodic functions, there exists a unique quadrature formula with optimal coefficients, and the optimal coefficients of which are a unique solution of the system of equations* (3.1)*, which admits the following representation:*

$$\mathring{A}_k = \frac{2i}{4\pi^2\omega^2+1}\cdot\frac{e^{2h}+1-2e^h\cos 2\pi\omega h}{e^h+1}\cdot\left[\frac{1}{2\pi\omega(e^h-1)} - \frac{\cot\pi\omega h}{e^h+1}\right]\cdot e^{2\pi i\omega hk},\ \ k=1,2,\dots,N. \qquad (3.3)$$

The proof of the theorem is the same as the proof of Theorem 3 in [5].
Thus, Problem 1.2 is solved.

4. A sharp upper bound for the error of the optimal quadrature formula

Now, we evaluate the error of the optimal quadrature formula of the form (1.2). To do this, we use the results of Theorems 2.2 and 3.1, that is, using the optimal coefficients $\overset{\circ}{A}_k$ determined by equation (3.3) when calculating the square of the norm for the error functional in Theorem 2.2, we obtain the following theorem

Theorem 4.1. *The norm of the error functional corresponding to the optimal quadrature formula of the form* (1.2)*, that is, the sharp upper bound of the error* (1.3)*, is expressed by the following expression:*

$$\left\|\overset{\circ}{\ell}\right\|^2 = \left\|\overset{\circ}{\ell}_2\right\|^2 - E^2(\omega, h),$$

where

$$\begin{aligned}\left\|\overset{\circ}{\ell}_2\right\|^2 &= \frac{1}{(2\pi\omega)^4 + (2\pi\omega)^2}\left[1 - \frac{2N}{(2\pi\omega)^4 + (2\pi\omega)^2}\right. \\ &\times \left.\left[\frac{h}{1-\cos 2\pi\omega h} - \frac{e^{2h}-1}{e^{2h} - 2e^h\cos 2\pi\omega h + 1}\right]^{-1}\right] \qquad (4.1)\end{aligned}$$

is given in Theorem 4 in [5] and

$$E^2(\omega, h) = \frac{1}{2h}\cdot\left(\frac{e^{2h}+1-2e^h\cos 2\pi\omega h}{(4\pi^2\omega^2+1)(e^h+1)}\right)^2\cdot\left(\frac{1}{\pi\omega(e^h-1)} - \frac{2\cot\pi\omega h}{e^h+1}\right)^2\cdot\left(\frac{e^h+1}{e^h-1} - \frac{2}{h}\right). \qquad (4.2)$$

To show that the function $E^2(\omega, h)$ is positive for arbitrary values of ω and h, we consider each multiplier in the formula separately. It is evident from formula (4.2) that the first factor $\frac{1}{2h}$ is positive for $h > 0$, and the second and third factors are non-negative as they are squares of real expressions. Therefore, the sign of $E^2(\omega, h)$ is determined by the last factor:

$$D(h) = \frac{e^h+1}{e^h-1} - \frac{2}{h}.$$

Using the identity $\frac{e^h+1}{e^h-1} = \coth(h/2)$, we can rewrite this term as:

$$D(h) = \coth\left(\frac{h}{2}\right) - \frac{1}{h/2}.$$

Using the inequality $\tanh(x) < x$ for $x > 0$, we have $\coth(x) = \frac{1}{\tanh(x)} > \frac{1}{x}$. Letting $x = h/2$, we obtain $\coth(h/2) > \frac{2}{h}$, which implies $D(h) > 0$. Thus, the function $E^2(\omega, h)$ is strictly positive for all $h > 0$ and ω.

Now we show numerically that the function $E^2(\omega, h)$ tends to zero as ω or N tends to infinity in Table 1.

	$\omega = 2$	$\omega = 11$	$\omega = 101$	$\omega = 1001$
$N = 10$	7.45e-8	1.08e-9	1.82e-13	1.92e-17
$N = 100$	9.23e-14	2.59e-12	2.09e-15	2.20e-19
$N = 1000$	9.25e-20	2.76e-18	2.21e-16	2.21e-21
$N = 10000$	9.26e-26	2.77e-24	2.33e-22	2.17e-20

Table 1. Value of the function $E^2(\omega, h)$ is positive and $E^2(\omega, h) \to 0$ as $\omega \to \infty$ or $N \to \infty$.

Theorem 4.1 shows that the absolute value of the error of the optimal quadrature formula constructed in this work is smaller than the absolute value of the error of the optimal quadrature formula in the work [5].

Conclusion

This work addresses the construction of an optimal quadrature formula involving derivatives, in the sense of Sard, for the approximation of strongly oscillatory integrals in the Hilbert space of complex-valued, periodic functions. To derive a sharp upper bound for the absolute error of the quadrature formula, the analytical expression for the norm of the associated error functional is obtained. This derivation begins with the identification of the extremal function corresponding to the error functional. Moreover, in this work, the optimal coefficients of the quadrature formula of the form (1.2) are derived. Additionally, the sharp upper bound for the error of the optimal quadrature formula is established, and it is demonstrated that this bound is smaller than the corresponding error obtained in the work [5].

References

[1] Xu Z., Milovanović G. V., Xiang S., Efficient computation of highly oscillatory integrals with Henkel kernel. Applied Mathematics and Computation, (2015) Vol. 261, P. 312–322.

[2] Milovanović G. V., Stanić M. P., Numerical Integration of Highly Oscillating Functions. Analytic Number Theory, Approximation Theory and Special Functions. Springer, (2014), P. 613–649.

[3] Olver S. Numerical Approximation of Highly Oscillatory Integrals. PhD thesis. University of Cambridge, 2008.

[4] Shadimetov K. M., Hayotov A. R., Nuraliev F. A., Optimal quadrature formulas of Euler–Maclaurin type. Applied Mathematics and Computation, (2016) Vol. 276, P. 340–355.

[5] Shadimetov K. M., Hayotov A. R., Khayriev U. N., Optimal quadrature formulas for approximating strongly oscillating integrals in the Hilbert space $\widetilde{W_2}^{(m,m-1)}$ of periodic function. Journal of Computational and Applied Mathematics, (2025) Vol. 423, P. 116–133.

[6] Shadimetov K. M., Hayotov A. R., Khayriev U. N., The Optimal Quadrature Formula for the Approximate Calculation of Fourier Coefficients in the Space $\widetilde{W_2}^{(2,1)}$ of Periodic Functions. Analysis, Approximation, Optimization: Computation and Applications: In Honor of Gradimir V. Milovanović on the Occasion of His 75th Birthday. P. 287–300.

[7] Shadimetov K. M., Hayotov A. R., Nuraliev F. A., Optimal interpolation formulas with derivative in the space $L_2^{(m)}(0,1)$. Filomat, (2019) Vol. 33, Iss. 17, P. 5661–5675.

[8] Hayotov A. R., Jeon S., Lee C.-O., Shadimetov K. M., Optimal quadrature formulas for non-periodic functions in Sobolev space and its application to CT image reconstruction. Filomat, (2021) Vol. 35, Iss. 12, P. 4177–4195.

[9] Hayotov A. R., Milovanović G. V., Shadimetov K. M., Optimal quadratures in the sense of Sard in a Hilbert space. Applied Mathematics and Computation, (2015) Vol. 259, P. 637–653.

[10] Hayotov A. R., Jeon S., Lee C. O., On an optimal quadrature formula for approximation of Fourier integrals in the space $L_2^{(1)}$. Journal of Computational and Applied Mathematics, (2020) Vol. 372, P. 112713.

[11] Hayotov A. R., Khayriev U. N., Optimal quadrature formulas in the space $\widetilde{W_2}^{(1,0)}$ of periodic functions. Uzbek Mathematical Journal, (2021) Vol. 65, Iss. 3, P. 93–100.

[12] Hayotov A. R., Khayriev U. N., An optimal quadrature formula for approximating Fourier integrals in a Hilbert space. AIP Conference Proceedings, (2024) Vol. 3004, Iss. 1, P. 060046.

[13] Hayotov A. R., Rasulov R. G., The order of convergence of an optimal quadrature formula with derivative in the space $W_2^{(2,1)}$. Filomat, (2020) Vol. 34, Iss. 11, P. 3835–3844.

[14] Boltaev N. D., Hayotov A. R., Milovanović G. V., Shadimetov K. M., Optimal quadrature formulas for Fourier coefficients in $W_2^{(m,m-1)}$ space. Journal of Applied Analysis and Computation, (2017) Vol. 7, P. 1233–1266.

[15] Jalolov O. I., Khayatov K. U., Algorithm for finding the norm of the error functional of a Hermite type cubature formula in the Sobolev space $\tilde{W}_2^{(m)}(T_n)$. AIP Conference Proceedings, (2024) Vol. 3004, P. 060038.

[16] Jalolov I. I., Jalolov O. I., An Algorithm for Constructing the Fourier Transform of the Function $V_m(x)$ to Determine a Discrete Analog of a Differential Operator. Russian Mathematics, (2024) Vol. 68, Iss. 12, P. 45–56.

[17] Hayotov A. R., Babaev S. S., Olimov N. N., An optimal interpolation formula of Hermite type in the Sobolev space. Filomat, (2024) Vol. 38, Iss. 23, P. 8305–8322.

[18] Hayotov A. R., Babaev S. S., Optimal quadrature formulas for computing of Fourier integrals in $W_2^{(m,m-1)}$ space. AIP Conference Proceedings, (2021) Vol. 2365, P. 020021.

[19] Babaev S. S., Consecutive optimization of the weighted quadrature formulas with derivative. Filomat, (2025) Vol. 39, Iss. 8, P. 2805–2815.

[20] Hayotov A. R., Khayriev U. N., Construction of an optimal quadrature formula in the Hilbert space of periodic functions. Lobachevskii Journal of Mathematics, (2022) Vol. 43, Iss. 11, P. 119–128.

[21] Akhmedov D. M., Atamuradova B. M., Construction of optimal quadrature formulas for Cauchy type singular integrals in the $W_2^{(1,0)}(0,1)$ space. Uzbek Mathematical Journal, (2022) Vol. 66, Iss. 2, P. 5–9.

[22] Akhmedov D. M., Optimal approximation of the Hadamard hypersingular integrals. Uzbek Mathematical Journal, (2023) Vol. 67, Iss. 3, P. 5–12.

[23] Akhmedov D. M., Hayotova S. A., Hayotov M. A., An optimal approximation formula for reconstruction of tomographic images of radial symmetric functions in the Sobolev space. Uzbek Mathematical Journal, (2024) Vol. 68, Iss. 1, P. 5–10.

[24] Hayotov A. R., Khayriev U. N., Optimal quadrature formulas for approximate calculation of integrals with exponential weight. Bulletin of the Institute of Mathematics, (2022) Vol. 5, Iss. 6, P. 14–22.

[25] Hayotov A. R., Khayriev U. N., Makhkamov D., Optimal quadrature formula for approximate calculation of integrals with exponential weight and its application. Bulletin of the Institute of Mathematics, (2021) Vol. 4, Iss. 2, P. 99–108.

[26] Khayriev U. N., Usmonova M. U., Turdieva G., An optimal quadrature formula in the space $\widetilde{W_2}^{(2,1)}$ of periodic complex-valued functions. AIP Conference Proceedings, (2024) Vol. 3004, Iss. 1, P. 060037.

[27] Sobolev S. L., Introduction to the Theory of Cubature Formulas. Nauka, Moscow, (1974).

[28] Hayotov A. R., Khayriev U. N., Farrukhbek A., Exponentially weighted optimal quadrature formula with derivative in the space $L_2^{(2)}$. AIP Conference Proceedings, (2023) Vol. 2781, P. 00050.

Khayriev U.N.,
Cyber University, Nurafshon, Uzbekistan
Bukhara State University, Bukhara, Uzbekistan
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences, Tashkent , Uzbekistan
Tashkent International University, Tashkent , Uzbekistan
e-mail: khayrievu@gmail.com, u.xayriyev@csu.uz

Nutfullaeva A.Kh.,
Bukhara State University, Bukhara, Uzbekistan
e-mail: khayrievu@gmail.com

Hamroeva Z.A.,
Bukhara State University, Bukhara, Uzbekistan
e-mail: khayrievu@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 145–150
DOI: 10.29229/uzmj.2026-2-17

Limit theorem for one scheme of particle evolution according to the critical Galton-Watson branching processes.

Khusanbaev Ya.M.

Abstract. We consider the critical Galton-Watson branching process a special form and possibly with the infinite variance the number of descendants of one particle. Local limit theorem for such process proved.

Keywords: Critical Galton-Watson branching process, generation function, local limit theorem
MSC (2020): 60J80, 60F05

1. Introduction

Let $\{Z(n),\ n \geq 0\}$ be a Galton-Watson branching process starting with a single particle ($Z(0) = 1$) in which particles leave immediate descendants according to probability law having a generating function

$$f(s) = \sum_{k=0}^{\infty} p_k s^k,\ \ 0 \leq s \leq 1,$$

where $p_k = P(Z(1) = k)$. The first work in which a conditional limit theorem for a subcritical Galton-Watson branching process was proved under the condition of non-degeneration of the process belong to A. M. Yaglom. Yaglom [1] proved for subcritical Galton-Watson branching processes that when $n \to \infty$ the conditional probabilities $P(Z(n) = k \,|Z(n) > 0)$ converge for all integers $k \geq 0$ and the limiting quantities form is probability distribution. An analogue of Yaglom's theorem for Markov branching processes with continuous time was proved by B.A. Sevastyanov [2]. The first work in which local limit theorems for Markov branching processes with continuous time were proved belongs to V.M. Zolotarev [3]. In it was found asymptotic behavior of the probability $P(Z(t) = k)$ as $t \to \infty$ for fixed k. Under the assumption of the existence of the fourth moment of the number of immediate descendants of one particle, V.P. Chistyakov [4] found an asymptotic behavior $P(Z(t) = k)$ as $t, k \to \infty$ for Markov branching processes with continuous time. S.V. Nagaev and R. Mukhamedkhanova [5] proved local limit theorem for critical Galton-Watson branching processes under the assumption of the existence of the fourth moment of the number of immediate descendants of one particle.
The local limit theorem for critical Galton-Watson branching processes was proved in the joint work of H.Kesten, P.Ney, F.Spitzer [6] under the condition $\sum\limits_{k=2}^{\infty} k^2 \ln k \cdot p_k < \infty$: if k and n tend to infinity so that their ratio remains bounded, then

$$\lim_{n\to\infty} \frac{\sigma^4 n^2}{4d} \exp\left\{\frac{2kd}{\sigma^2 n}\right\} P(Z(n) = kd) = 1, \tag{1.1}$$

where $\sigma^2 = f''(1),\ d = g.c.d.\{k : p_k > 0\}$.

Despite the assertion of the authors of [6] that for the validity of the local limit theorem for critical Galton-Watson branching processes it is sufficient that the variance of the number of descendants of one particle be finite, no proof of this assertion was known for more than forty years. The final version of the proof of the local limit theorem for critical Galton-Watson branching processes under the condition of finite variance of the number of immediate descendants belongs to S.V. Nagaev and V. Vakhtel [7]: let $\sigma^2 < \infty$. Then for all $k \leq n$

$$\lim_{n\to\infty} \frac{\sigma^4 n^2}{4d} \left(1 + \frac{2d}{\sigma^2 n}\right)^{k+1} P(Z(n) = kd) = 1. \tag{1.2}$$

We also note that the question of the validity of the local limit theorem for critical Bellman-Harris processes was investigated by V.A. Vatutin [8], V.A. Topchiy [9] proved an analogue of (1.1) for critical

Bellman-Harris processes with discrete time. We also note the works of K.B. Athreya, P. Ney [10], S. Dubic and E. Seneta [11], H. Imai [12] in which the validity of the local limit theorem for supercritical Galton-Watson branching processes was investigated.

Let the generating function $f(s)$ satisfy the condition

$$f(s) = s + (1-s)^{1+\alpha} L(1-s), \ 0 \le s \le 1, \tag{1.3}$$

where $\alpha \in (0,1]$, $L(x)$- slowly varying function as $x \to +0$.

Let $f_0(s) = s$ and we defined the iteration of the function $f(s)$ from relation $f_n(s) = f(f_{n-1}(s)), \ n = 1,2,...$ It is well known (see [13]) that when the condition (1.3) hold, then

$$(1 - f_n(0))^{\alpha} L(1 - f_n(0)) \sim \frac{1}{\alpha n} \tag{1.4}$$

as $n \to \infty$ and

$$Q(n) := P(Z(n) > 0) = 1 - f_n(0) \sim \frac{1}{[\alpha n L(1 - f_n(0))]^{1/\alpha}} \tag{1.5}$$

as $n \to \infty$. Moreover, for anyone $x \ge 0$

$$F(x) = \lim_{n\to\infty} P(Q(n) Z(n) \le x | Z(n) > 0) =$$
$$= \begin{cases} 1 - e^{-x}, & as \ \alpha = 1, \\ 1 - \frac{\alpha}{\Gamma(1/\alpha)} \int\limits_0^{\infty} e^{-(x/u)^{\alpha}} p(u, \alpha, 1)\, du, & as \ 0 < \alpha < 1, \end{cases}$$

where $p(u, \alpha, 1)$ is the density of one-sided stable law with Laplace transform

$$\int_0^{\infty} e^{-\lambda u} p(u, \alpha, 1)\, du = e^{-\lambda^{\alpha}}, \ \lambda \ge 0.$$

It is known [1] that the Laplace transform $\psi(\lambda)$ corresponding to the distribution $F(x)$ has the form

$$\psi(\lambda) = 1 - \lambda (1 + \lambda^{\alpha})^{-1/\alpha}.$$

We note that in the case $\alpha = 0$ of the well-known result of S.V. Nagaev and V. Vakhtel [7] that the conditional distribution of the correspondingly transformed $Z(n)$ under the condition of non-degeneration $(Z(n) > 0)$ by the moment n weakly converges to the exponential law.

Slack [13] proved that if (1.3) hold, then for any (fixed) $i, j \ge 1$ there are limit

$$\lim_{n\to\infty} L^{1/\alpha}(1 - f_n(0)) (\alpha n)^{1+1/\alpha} P_n(i,j) = i\pi(j), \tag{1.6}$$

at that

$$\sum_{i=1}^{\infty} \pi(i) P_1(i,j) = \pi(j), \ j \ge 1, \ \sum_{i=1}^{\infty} \pi(i) p_0^i = 1 \tag{1.7}$$

and as $J \to \infty$

$$\pi(1) + \pi(2) + ... + \pi(J) \sim \frac{1}{\alpha(1+\alpha)} J^{\alpha} \frac{1}{L\left(\frac{1}{J}\right)}, \tag{1.8}$$

where $P_n(i,j)$ is probability transition of the chain $\{Z(n), \ n \ge 0\}$ from state i to state j for n-step.

We note that relations (1.6)-(1.8) are proved in [6] for the case $\alpha = 1, \ \sigma^2 < \infty$.
However the asymptotic of the probability $P_n(i,j)$ for $n, j \to \infty$ in the case when (1.3) and $\sigma^2 = \infty$ unknown.

Let condition A be satisfied: the probability generating function $f(s)$ has the form

$$f(s) = 1 - \left[C\alpha + (1-s)^{-\alpha}\right]^{-1/\alpha}, \ s \in [0,1), \ f(1) = 1, \tag{1.9}$$

where $\alpha \in (0,1]$, C is positive constant number.

It is easy to see that

$$f(s) = s + (1-s)^{1+\alpha} L^*(1-s)$$

as $s \to 1$, here $L^*(x) \to C$ as $x \to 0$, $f'(1) = 1$ and $f''(1) = +\infty$ for $\alpha \in (0,1)$.

The purpose of this paper is to establish the asymptotic of the transition probabilities $P_n(1,k)$ of the critical Galton-Watson branching process $\{Z(n), \ n \ge 0\}$under condition A in the case when $n, k \to \infty, \ k = o\left(n^{1/(2-\alpha)}\right)$.

2. Formulation of the result.

Theorem 2.1. *Let the probability generating function $f(s)$ satisfy condition A. Then*
1. for any fixed integer $k \geq 1$

$$C^{1/\alpha}\left[\alpha n\right]^{1+1/\alpha} P_n(1,k) \to \pi_k := \frac{1}{Ck}\prod_{i=1}^{k-1}\left(1+\frac{\alpha}{i}\right) \tag{2.1}$$

as $n \to \infty$;
2. as $n \to \infty,\ k \to \infty,\ k = o\left(n^{1/(2-\alpha)}\right)$

$$(C\alpha n)^{1/\alpha+1}\,\Gamma(\alpha+1)\,k^{1-\alpha}P_n(1,k) \to 1, \tag{2.2}$$

3. as $n \to \infty,\ k \to \infty,\ k = o(n)$

$$\pi_k - \frac{f_n^{(k)}(0)}{Ck!f_n'(0)} = \mathrm{O}\left(\frac{k}{n}\right),\ \ \pi_k \sim \frac{1}{C\Gamma(1+\alpha)\,k^{1-\alpha}}. \tag{2.3}$$

3. Auxiliary concepts and results.

Let $N = \{1,2,...\}$ and $N_0 = N \cup \{0\}$. For each pair of integers $k \leq J$, we introduce the set

$$D(J,k) := \left\{i_1 + i_2 + ... + i_{J-k+1} = k,\ i_1 + 2i_2 + ... + (J-k+1)\,i_{J-k+1} = J\right\}.$$

Bell polynomials of the second type are functions $B_{J,k}(x_1,...,x_{J-k+1})$, $k \leq J$ determined from the relation

$$\frac{1}{k!}\left(\sum_{j=1}^{\infty} x_j \frac{s^j}{j!}\right)^k = \sum_{J=k}^{\infty} B_{J,k}(x_1,...,x_{J-k+1})\frac{s^J}{J!} \tag{3.1}$$

and they can be calculated using the formula

$$B_{J,k}(x_1,...,x_{J-k+1}) = \sum_{D(J,k)} \frac{J!}{i_1!\cdot...\cdot i_{J-k+1}!}\prod_{r=1}^{J-k+1}\left(\frac{x_r}{r!}\right)^{i_r}.$$

Note that due to (3.1)

$$B_{J,1}(x_1,...,x_J) = \sum_{D(J,k)} \frac{J!}{i_1!\cdot...\cdot i_J!}\prod_{r=1}^{J}\left(\frac{x_r}{r!}\right)^{i_r} = x_J,$$

$$B_{J,k}(\gamma x_1,...,\gamma x_{J-k+1}) = \gamma^k B_{J,k}(x_1,...,x_{J-k+1}). \tag{3.2}$$

It is known ([14], relations (1.7)) that

$$B_{n,k}(1!,2!,...,(n-k+1)!) = C_{n-1}^{k-1}\frac{n!}{k!}, \tag{3.3}$$

here

$$C_n^k = \frac{n!}{k!\,(n-k)!}.$$

We will use Bell polynomials of the second kind to find derivatives of a superposition function $F(G(x))$ from differentiable functions $F(x)$ and $G(x)$ using the Faa-De Bruno formula:

$$\frac{d^k}{dx^k}F(G(x)) = F'(G(x))\,G^{(k)}(x) + \sum_{J=2}^{k} F^{(J)}(G(x))\,B_{k,J}\left(G'(x),...,G^{(k-J+1)}(x)\right), \tag{3.4}$$

where the second term is considered equal to zero if $J = 1$.

4. Proof of the theorem.

It is easy to see that

$$f_n(s) = 1 - \left[C\alpha n + (1-s)^{-\alpha}\right]^{-1/\alpha}.$$

We denote

$$F(x) = 1 - x^{-1/\alpha},\ G(x) = C\alpha n + (1-x)^{-\alpha}.$$

We have

$$F^{(k)}(x) = (-1)^{k+1}\frac{1}{\alpha}\left(\frac{1}{\alpha}+1\right)...\left(\frac{1}{\alpha}+k-1\right)x^{-1/\alpha-k},$$

$$G^{(k)}(x) = \alpha(\alpha+1)...(\alpha+k-1)(1-x)^{-\alpha-k}.$$

According (3.4) we have

$$f_n^{(k)}(s) = [\,\mathrm{F}(\mathrm{G}(s))]^{(k)} = \frac{1}{\alpha}\left[C\alpha n + (1-s)^{-\alpha}\right]^{-1/\alpha-1}\prod_{i=0}^{k-1}(\alpha+i)(1-s)^{-\alpha-k} +$$

$$+\sum_{j=2}^{k}(-1)^{j+1}\prod_{i=0}^{j-1}\left(\frac{1}{\alpha}+i\right)\left[C\alpha n + (1-s)^{-\alpha}\right]^{-\frac{1}{\alpha}-j}\times$$

$$\times B_{k,j}\left(\alpha(1-s)^{-\alpha-1}, \alpha(\alpha+1)(1-s)^{-\alpha-2}, ..., \prod_{i=0}^{k-j}(\alpha+i)(1-s)^{-\alpha-k+j-1}\right).$$

From here, considering the relation (3.1) and (3.2), we obtain

$$f_n^{(k)}(0) = \frac{1}{\alpha}[C\alpha n+1]^{-1/\alpha-1}\prod_{i=0}^{k-1}(\alpha+i) +$$

$$+\sum_{j=2}^{k}(-1)^{j+1}\prod_{i=0}^{j-1}\left(\frac{1}{\alpha}+i\right)[C\alpha n+1]^{-\frac{1}{\alpha}-j}B_{k,j}\left(\alpha, \alpha(\alpha+1), ..., \alpha\prod_{i=1}^{k-j}(\alpha+i)\right) =$$

$$= \frac{1}{\alpha}[C\alpha n+1]^{-1/\alpha-1}\prod_{i=0}^{k-1}(\alpha+i) + \Delta_{k,n}, \tag{4.1}$$

where

$$\Delta_{k,n} = \sum_{j=2}^{k}(-1)^{j+1}\prod_{i=0}^{j-1}\left(\frac{1}{\alpha}+i\right)[C\alpha n+1]^{-\frac{1}{\alpha}-j}\alpha^j B_{k,j}\left(1, (\alpha+1), ..., \prod_{i=1}^{k-j}(\alpha+i)\right). \tag{4.2}$$

From (3.1) follows

$$B_{k,j}\left(1, (\alpha+1), ..., \prod_{i=1}^{k-j}(\alpha+i)\right) \le B_{k,j}(1!, 2!, ..., (k-j+1)!). \tag{4.3}$$

Further, relations (4.2)-(4.3) and (3.3) lead us to the estimate

$$|\Delta_{k,n}| \le [C\alpha n+1]^{-1/\alpha-2}\sum_{j=2}^{k}\prod_{i=0}^{j-1}\left(\frac{1}{\alpha}+i\right)\alpha^j C_{k-1}^{j-1}\frac{k!}{j!}n^{-j+2} \le \tag{4.4}$$

$$\le [C\alpha n+1]^{-1/\alpha-2}\,kk!\sum_{j=2}^{k}\frac{1}{j!}\alpha^{j-1}\prod_{i=1}^{j-1}\left(1+\frac{1}{i\alpha}\right).$$

The series $\sum_{j=2}^{\infty}\frac{1}{j!}\alpha^j\prod_{i=1}^{j-1}\left(1+\frac{1}{i\alpha}\right)$ is converges. Now from (4.1) and (4.4) we conclude that

$$P_n(1,k)=\frac{1}{k!}f_n^{(k)}(0)=\frac{1}{k}\left[C\alpha n+1\right]^{-1/\alpha-1}\prod_{i=1}^{k-1}\left(1+\frac{\alpha}{i}\right)+\delta_{k,n}, \tag{4.5}$$

here

$$|\delta_{k,n}|\le\left|\frac{1}{k!}\Delta_{k,n}\right|\le\left[C\alpha n+1\right]^{-1/\alpha-2}k\sum_{j=2}^{\infty}\frac{1}{j!}\alpha^{j-1}\prod_{i=1}^{j-1}\left(1+\frac{1}{i\alpha}\right),$$

from which the validity of (2.1) follows. Now we write (4.5) in the form

$$k\left[C\alpha n+1\right]^{1/\alpha+1}\left[\prod_{i=1}^{k-1}\left(1+\frac{\alpha}{i}\right)\right]^{-1}P_n(1,k)=1+\tau_{k,n}, \tag{4.6}$$

here

$$|\tau_{k,n}|\le\left[C\alpha n+1\right]^{-1}k^2\left[\prod_{i=1}^{k}\left(1+\frac{\alpha}{i}\right)\right]^{-1}\sum_{j=2}^{\infty}\frac{1}{j!}\alpha^{j-1}\prod_{i=1}^{j-1}\left(1+\frac{1}{i\alpha}\right)$$

Further, according to the Weierstrass's formula

$$\Gamma(\alpha)=\frac{e^{-\alpha\gamma}}{\alpha}\prod_{i=1}^{\infty}\left(1+\frac{\alpha}{i}\right)^{-1}e^{\alpha/i},$$

here γ is the Euler-Mascheroni's constant. Hence considering the main property Γ-function ($\alpha\Gamma(\alpha)=\Gamma(\alpha+1)$) and Euler's formula for the harmonic series we find for large$k$

$$\prod_{i=1}^{k}\left(1+\frac{\alpha}{i}\right)=\frac{1}{\alpha e^{\alpha\gamma}\Gamma(\alpha)}\prod_{i=1}^{k}e^{\alpha/i}\left(1+o(1)\right)=\frac{1}{\alpha e^{\alpha\gamma}\Gamma(\alpha)}e^{\alpha\sum_{i=1}^{k}\frac{1}{i}}\left(1+o(1)\right)=$$

$$=\frac{1}{\alpha e^{\alpha\gamma}\Gamma(\alpha)}e^{\alpha(\gamma+\ln k)}\left(1+o(1)\right)=\frac{k^\alpha}{\alpha\Gamma(\alpha)}\left(1+o(1)\right)=\frac{k^\alpha}{\Gamma(\alpha+1)}\left(1+o(1)\right). \tag{4.7}$$

From (4.6)-(4.7) as $n\to\infty,\ k\to\infty,\ k=o\left(n^{\frac{1}{2-\alpha}}\right)$ we conclude

$$(C\alpha n)^{1/\alpha+1}\Gamma(\alpha+1)k^{1-\alpha}P_n(1,k)=1+\mathrm{O}\left(\frac{k^{2-\alpha}}{n}\right)\to 1,$$

which proves the validity of (2.2). Now we will show the validity of (2.3). We put

$$q_{k,n}=\frac{P_n(1,k)}{CP_n(1,1)}=\frac{f_n^{(k)}(0)}{Ck!f_n'(0)},\ k=1,2,...$$

From (4.1) we have

$$f_n'(0)=\left[C\alpha n+1\right]^{-1/\alpha-1}. \tag{4.8}$$

From (4.5) and (4.8) we conclude that for any fixed $k\ge 1$ the sequence $q_{k,n},\ n\ge 1$ converges to a finite limit

$$\lim_{n\to\infty}q_{k,n}=\pi_k=\frac{1}{Ck}\prod_{i=1}^{k-1}\left(1+\frac{\alpha}{i}\right). \tag{4.9}$$

Further, by virtue of (4.1), (4.4) and (4.8) we have

$$|q_{k,n+1}-q_{k,n}|\le K\frac{k}{Cn^2}\sum_{j=2}^{k}\prod_{i=1}^{j-1}\left(1+\frac{1}{i\alpha}\right)\cdot\frac{1}{(j-1)!}$$

where K is positive constant depending on α.
Conseq uently,

$$|\pi_k - q_{k,n}| = \left|\sum_{r=n}^{\infty}(q_{k,r+1} - q_{k,r})\right| \leq K\frac{k}{Cn}\sum_{j=2}^{\infty}\prod_{i=1}^{j-1}\left(1+\frac{1}{i\alpha}\right)\cdot\frac{1}{(j-1)!}.$$

It follows that for all $k = o(n)$ as $n \to \infty$

$$\sup_{k\leq n, k=o(n)} |\pi_k - q_{k,n}| \to 0.$$

From the relation (4.7) and (4.9) we concluded that

$$\pi_k \sim \frac{1}{\Gamma(1+\alpha)\, k^{1-\alpha}}$$

as $k \to \infty$. By (4.9) it is easy to see that the sequence $\{\pi_k, k \in N\}$ is monotonically decreasing.
The proof of the theorem is complete.

References

[1] Yaglom A. M., Certain limit theorems of the theory of branching random processes. Doklady Akad. Nauk. SSSR, (1947) Vol. 56, P. 347–349.

[2] Sevastyanov B. A., The theory of branching random processes. Uspekhi Matematicheskikh Nauk, (1951) Vol. 6, Iss. 6, P. 47–99.

[3] Zolotarev V. M., More exact statements of several theorems in the theory of branching processes. Theory of Probability and Its Applications, (1957) Vol. 2, Iss. 2, P. 245–253.

[4] Chistyakov V. P., Lokal limit theorems in the theory of branching random processes. Theory of Probability and Its Applications, (1957) Vol. 2, Iss. 3, P. 345–363.

[5] Nagaev S. V., Mukhamedkhanova R., Some limit theorems from the theory branching processes. Limit Theorems and Statistical Inferences, (1966), P. 90–112.

[6] Kesten H., Ney P. E., Spitzer F. L., The Galton-Watson process with mean one and finite variance. Theory of Probability and Its Applications, (1966) Vol. 11, Iss. 4, P. 579–611.

[7] Nagaev S. V., Vakhtel V. I., On the local limit theorem for a critical Galton-Watson process. Theory of Probability and Its Applications, (2006) Vol. 50, Iss. 3, P. 400–419.

[8] Vatutin V. A., Local limit theorem for critical Bellman-Harris branching processes. Proceedings of the Steklov Mathematical Institute of the USSR Academy of Sciences, (1981) Vol. 158, P. 9–30.

[9] Topchiy V. A., Local limit theorem for critical Bellman-Harris processes with discrete time. Limit theorems of probability theory and related questions. Nauka, (1982), P. 97–122.

[10] Athreya K. B., Ney P., The local limit theorem and some related aspects of super-critical branching processes. Transactions of the American Mathematical Society, (1970) Vol. 152,

[11] Dubuc S., Seneta E., The Local Limit Theorem for the Galton-Watson Process. The Annals of Probability, (1976) Vol. 4, Iss. 3,

[12] Imai H., Notes on a local limit theorem for discrete time Galton-Watson branching processes. Annals of the Institute of Statistical Mathematics, (1968) Vol. 20, P. 391–410.

[13] Slack R., A branching process with mean one and possibly infinite variance. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, (1968) Vol. 9, P. 139–145.

[14] Qi F., Niu D.-W., Lim D., Yao Y.-H., Special values of the Bell polynomials of the second kind for some sequences and functions. J. Math. Anal. Appl. (2020) Vol. 491,

Khusanbaev Ya.M.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences, Tashkent , Uzbekistan
e-mail: ykhusanbaev@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 151–158
DOI: 10.29229/uzmj.2026-2-18

On fixed point results in a new type of nonlinear contraction in $\mathfrak{b}$-metric spaces with applications to differential equations

Mahdi H. S., Rasouli S. H., Mohammadzadeh B.

Abstract. We present fixed point theorem for a new type of nonlinear contractive mappings in $\mathfrak{b}$-metric spaces. Also, we present examples to illustrate the validity of the results obtained in the paper. In addition, by using our results, we obtain the existence and uniqueness of solution to some ordinary differential equations with initial value conditions.

Keywords: Nonlinear contraction; $\mathfrak{b}$-metric spaces; Fixed point results.
MSC (2020): 47H10, 54H25

1. Introduction

Fixed point theory is one of the most well-known and established theories in mathematics and has a variety of applications. In this theory, contraction is one of the main tools to prove the existence and uniqueness of a fixed point. The formal form of the metric fixed point theory appeared by the pioneer and art work of Banach results in 1922 : Let (X, d) be a complete metric space, $T : X \to X$ be a contraction, there exists a unique fixed point $x_0 \in X$ of T. Its broad applicability has led to numerous extensions and generalizations over the years. One such generalization is the concept of a $\mathfrak{b}$-metric space, introduced by Bakhtin [1] and further studied by Czerwik [2, 3].

A $\mathfrak{b}$-metric space differs from a metric space in that the triangle inequality is relaxed. The relaxation of the triangle inequality makes $\mathfrak{b}$-metric spaces suitable for modeling certain situations where the conventional triangle inequality is not satisfied, such as those arising in computer science [4] and biological modeling [5]. Since their inception, $\mathfrak{b}$-metric spaces have attracted significant attention, and numerous fixed point theorems have been established in this setting (see [6, 7, 8, 9, 10, 11]).

Recently, the focus has shifted towards exploring nonlinear contractions, which provide a more flexible framework than linear contractions. These contractions involve functions or inequalities that incorporate the distance between points in a more complex manner, often utilizing auxiliary functions and parameters (see [12, 13]).

This paper aims to contribute to the ongoing development of fixed point theory in $\mathfrak{b}$-metric spaces by presenting a fixed point theorem for new type of nonlinear contractive mappings. We then demonstrate the practical relevance of these results by applying them to prove the existence and uniqueness of solutions to ordinary differential equations. This connection between abstract fixed point theorems and concrete problems in differential equations highlights the power and versatility of fixed point theory.

2. Preliminaries

In this section, we present a few key terms and definitions that will be used in our discussion.
Definition 2.1 [1]. Let $\mathfrak{X}$ be a nonempty set and $s \geq 1$ be a real number. A function $\mathfrak{d} : \mathfrak{X} \times \mathfrak{X} \to [0, \infty)$ is called a $\mathfrak{b}$-metric if for all $x, y, z \in \mathfrak{X}$, the following conditions are satisfied:
($b1$) $\mathfrak{d}(x, y) = 0$ if and only if $x = y$;
($b2$) $\mathfrak{d}(x, y) = \mathfrak{d}(y, x)$;

$(b3)$ $\mathfrak{d}(x,z) \leq s[\mathfrak{d}(x,y) + \mathfrak{d}(y,z)]$.

The pair $(\mathfrak{X}, \mathfrak{d})$ is called a $\mathfrak{b}$-metric space with parameter s.

Example 2.2 ([14]). Let $(\mathfrak{X}, d)$ be a metric space and let the mapping $\mathfrak{d} : \mathfrak{X} \times \mathfrak{X} \to [0, \infty)$ be defined by

$$\mathfrak{d}(p,q) := (d(p,q))^r, \text{for all } p, q \in \mathfrak{X},$$

where $r > 1$ is a fixed real number. Then $(\mathfrak{X}, \mathfrak{d})$ is a $\mathfrak{b}$-metric space with $s = 2^{r-1}$.

Remark 2.3. It should be noted that, the class of $\mathfrak{b}$-metric spaces is effectively larger than that of metric spaces, since a $\mathfrak{b}$-metric is a metric when $s = 1$. Also in general a $\mathfrak{b}$-metric need not necessarily be a metric. For example, if $\mathfrak{X} = \mathbb{R}$ is the set of real numbers and $\mathfrak{d}(x,y) = |x - y|$ is the usual Euclidean metric, then $\eta(x,y) = (x-y)^2$ is a $\mathfrak{b}$-metric on $\mathbb{R}$ with $s = 2$, but is not a metric on $\mathbb{R}$.

Now, we present the concepts of convergence, Cauchy sequence and completeness in b-metric spaces.

Definition 2.4 ([15]). Let $(\mathfrak{X}, \mathfrak{d})$ be a $\mathfrak{b}$-metric space. Then a sequence $\{x_n\}$ in $\mathfrak{X}$ is

1. *converges* to $x \in \mathfrak{X}$ if and only if $\lim_{n\to\infty} \mathfrak{d}(x_n, x) = 0$;
2. *Cauchy sequence* if $\lim_{n,m\to\infty} \mathfrak{d}(x_n, x_m) = 0$.

Definition 2.5 ([15]). The $\mathfrak{b}$-metric space $(\mathfrak{X}, \mathfrak{d})$ is complete if every Cauchy sequence in $\mathfrak{X}$ converges to a point in $\mathfrak{X}$.

Now, we recall an existing fixed point theorem in $\mathfrak{b}$-metric spaces that will be useful for comparison and motivation for our new results.

Theorem 2.6 ([2].) Let $(\mathfrak{X}, \mathfrak{d})$ be a complete $\mathfrak{b}$-metric space and let $\mathfrak{T} : \mathfrak{X} \to \mathfrak{X}$ be a mapping such that for all $x, y \in \mathfrak{X}$,

$$\mathfrak{d}(\mathfrak{T}x, \mathfrak{T}y) \leq k\mathfrak{d}(x,y),$$

where $0 \leq k < 1/s$. Then $\mathfrak{T}$ has a unique fixed point in $\mathfrak{X}$.

3. Main results

In this section, we introduce our new type of nonlinear contraction mapping and establish fixed point theorems for such mappings in complete $\mathfrak{b}$-metric spaces. At first, we prove a fixed point theorem for generalized contraction mappings.

Theorem 3.2. Let $(\mathfrak{X}, \mathfrak{d})$ be a complete $\mathfrak{b}$-metric space and let $\mathfrak{T} : \mathfrak{X} \to \mathfrak{X}$ be a mapping such that for all $x, y \in \mathfrak{X}$,

$$\mathfrak{d}(\mathfrak{T}x, \mathfrak{T}y) \leq a\mathfrak{d}(x,y) + b[\mathfrak{d}(x, \mathfrak{T}x) + \mathfrak{d}(y, \mathfrak{T}y)],$$

where $a, b \geq 0$ and $s(a + 2b) < 1$. Then $\mathfrak{T}$ has a unique fixed point in $\mathfrak{X}$.

Proof. Let $x_0 \in \mathfrak{X}$ be an arbitrary point. Define a sequence $\{x_n\}$ in $\mathfrak{X}$ by $x_{n+1} = \mathfrak{T}x_n$ for all $n \geq 0$. Then, we have

$$\begin{aligned} \mathfrak{d}(x_{n+1}, x_{n+2}) &= \mathfrak{d}(\mathfrak{T}x_n, \mathfrak{T}x_{n+1}) \\ &\leq a\mathfrak{d}(x_n, x_{n+1}) + b[\mathfrak{d}(x_n, \mathfrak{T}x_n) + \mathfrak{d}(x_{n+1}, \mathfrak{T}x_{n+1})] \\ &= a\mathfrak{d}(x_n, x_{n+1}) + b[\mathfrak{d}(x_n, x_{n+1}) + \mathfrak{d}(x_{n+1}, x_{n+2})]. \end{aligned}$$

Rearranging, we get

$$\mathfrak{d}(x_{n+1}, x_{n+2}) \leq \frac{a+b}{1-b}\mathfrak{d}(x_n, x_{n+1}) = k\mathfrak{d}(x_n, x_{n+1}),$$

where $k = a + b/1 - b$. Since $s(a + 2b) < 1$, we have $s(a + b) < 1 - b$, which implies $k < 1/s$. Now, by induction, we have

$$\mathfrak{d}(x_{n+1}, x_{n+2}) \leq k^{n+1}\mathfrak{d}(x_0, x_1).$$

For $m > n$, we have

$$\begin{aligned}
\mathfrak{d}(x_n, x_m) &\leq s[\mathfrak{d}(x_n, x_{n+1}) + \mathfrak{d}(x_{n+1}, x_m)] \\
&\leq s\mathfrak{d}(x_n, x_{n+1}) + s^2\mathfrak{d}(x_{n+1}, x_{n+2}) + ... + s^{m-n}\mathfrak{d}(x_{m-1}, x_m) \\
&\leq sk^n\mathfrak{d}(x_0, x_1) + s^2k^{n+1}\mathfrak{d}(x_0, x_1) + ... + s^{m-n}k^{m-1}\mathfrak{d}(x_0, x_1) \\
&= sk^n[1 + sk + (sk)^2 + ... + (sk)^{m-n-1}]\mathfrak{d}(x_0, x_1) \\
&\leq \frac{sk^n}{1 - sk}\mathfrak{d}(x_0, x_1).
\end{aligned}$$

Since $k < 1/s$, we have $sk < 1$. Therefore, $\lim_{n,m\to\infty} \mathfrak{d}(x_n, x_m) = 0$, which implies that $\{x_n\}$ is a Cauchy sequence in $\mathfrak{X}$. Since $(\mathfrak{X}, \mathfrak{d})$ is complete, there exists $x^* \in \mathfrak{X}$ such that $\lim_{n\to\infty} x_n = x^*$. Now, we will show that x^* is a fixed point of $\mathfrak{T}$.

$$\begin{aligned}
\mathfrak{d}(\mathfrak{T}x^*, x_{n+1}) &= \mathfrak{d}(\mathfrak{T}x^*, \mathfrak{T}x_n) \leq a\mathfrak{d}(x^*, x_n) + b[\mathfrak{d}(x^*, \mathfrak{T}x^*) + \mathfrak{d}(x_n, \mathfrak{T}x_n)] \\
&= a\mathfrak{d}(x^*, x_n) + b[\mathfrak{d}(x^*, \mathfrak{T}x^*) + \mathfrak{d}(x_n, x_{n+1})].
\end{aligned}$$

Taking the limit as $n \to \infty$, we get $\mathfrak{d}(\mathfrak{T}x^*, x^*) \leq b\mathfrak{d}(x^*, \mathfrak{T}x^*)$. Therefore, $(1 - b)\mathfrak{d}(\mathfrak{T}x^*, x^*) \leq 0$. Since $b < 1/2s \leq 1$, we have $1 - b > 0$ and therefore, $d(\mathfrak{T}x^*, x^*) = 0$, which implies $\mathfrak{T}x^* = x^*$. Thus, x^* is a fixed point of $\mathfrak{T}$.

To prove uniqueness, suppose y^* is another fixed point of $\mathfrak{T}$ such that $y^* \neq x^*$. Then,

$$\mathfrak{d}(x^*, y^*) = \mathfrak{d}(\mathfrak{T}x^*, \mathfrak{T}y^*) \leq a\mathfrak{d}(x^*, y^*) + b[\mathfrak{d}(x^*, \mathfrak{T}x^*) + \mathfrak{d}(y^*, \mathfrak{T}y^*)] = a\mathfrak{d}(x^*, y^*).$$

Therefore, $(1 - a)\mathfrak{d}(x^*, y^*) \leq 0$. Since $a < 1/s \leq 1$, we have $1 - a > 0$. Therefore, $\mathfrak{d}(x^*, y^*) = 0$, which implies $x^* = y^*$. Hence, the fixed point is unique. This completes the proof. □

Definition 3.2. Let $(\mathfrak{X}, \mathfrak{d})$ be a $\mathfrak{b}$-metric space and $\mathfrak{T} : \mathfrak{X} \to \mathfrak{X}$ be a mapping. We say that $\mathfrak{T}$ is a φ-nonlinear contraction if there exists a function $\varphi : [0, \infty) \to [0, \infty)$ satisfying $\varphi(t) < t$ for all $t > 0$ and $\varphi(0) = 0$, and a constant $k \in (0, 1)$ such that

$$\mathfrak{d}(\mathfrak{T}x, \mathfrak{T}y) \leq k\varphi(\mathfrak{M}(x, y)), \forall x, y \in X, \tag{3.1}$$

where

$$\mathfrak{M}(x, y) = \max\left\{\mathfrak{d}(x, y), \mathfrak{d}(x, \mathfrak{T}x), \mathfrak{d}(y, \mathfrak{T}y), \frac{\mathfrak{d}(x, \mathfrak{T}y) + \mathfrak{d}(y, \mathfrak{T}x)}{2s}\right\}.$$

Theorem 3.3. Let $(\mathfrak{X}, d)$ be a complete $\mathfrak{b}$-metric space and let $\mathfrak{T} : \mathfrak{X} \to \mathfrak{X}$ be a φ-nonlinear contraction mapping such that $\max\{k, \frac{k}{2-k}\} < \frac{1}{s}$. Then $\mathfrak{T}$ has a unique fixed point in $\mathfrak{X}$.

Proof. Let x_0 be an arbitrary point in $\mathfrak{X}$. We define a sequence $\{x_n\}$ in $\mathfrak{X}$ by $x_{n+1} = \mathfrak{T}x_n$ for all $n \geq 0$. If for some N, $x_N = x_{N+1}$, then $x_N = \mathfrak{T}x_N$, which means x_N is a fixed point. In this case, the theorem holds. Assume that $x_n \neq x_{n+1}$ for all $n \geq 0$, so $\mathfrak{d}(x_n, x_{n+1}) > 0$ for all n. Consider $\mathfrak{d}(x_n, x_{n+1}) = \mathfrak{d}(\mathfrak{T}x_{n-1}, \mathfrak{T}x_n)$ for $n \geq 1$. Applying the contraction condition (1), we have

$$\mathfrak{d}(x_n, x_{n+1}) \leq k\varphi(\mathfrak{M}(x_{n-1}, x_n),)$$

where

$$\mathfrak{M}(x_{n-1}, x_n) = \max\left\{\mathfrak{d}(x_{n-1}, x_n), \mathfrak{d}(x_{n-1}, \mathfrak{T}x_{n-1}), \mathfrak{d}(x_n, \mathfrak{T}x_n), \frac{\mathfrak{d}(x_{n-1}, \mathfrak{T}x_n) + \mathfrak{d}(x_n, \mathfrak{T}x_{n-1})}{2s}\right\}.$$

Substituting $x_n = \mathfrak{T}x_{n-1}$ and $x_{n+1} = \mathfrak{T}x_n$:

$$\mathfrak{M}(x_{n-1}, x_n) = \max\left\{\mathfrak{d}(x_{n-1}, x_n), \mathfrak{d}(x_{n-1}, x_n), \mathfrak{d}(x_n, x_{n+1}), \frac{\mathfrak{d}(x_{n-1}, x_{n+1}) + \mathfrak{d}(x_n, x_n)}{2s}\right\}.$$

Since $\mathfrak{d}(x_n, x_n) = 0$, this simplifies to:

$$\mathfrak{M}(x_{n-1}, x_n) = \max\left\{\mathfrak{d}(x_{n-1}, x_n), \mathfrak{d}(x_n, x_{n+1}), \frac{\mathfrak{d}(x_{n-1}, x_{n+1})}{2s}\right\}.$$

Now, we analyze the possible cases for $\mathfrak{M}(x_{n-1}, x_n)$:

Case A. Suppose $\mathfrak{M}(x_{n-1}, x_n) = \mathfrak{d}(x_n, x_{n+1})$. Then

$$\mathfrak{d}(x_n, x_{n+1}) \leq k\varphi(\mathfrak{d}(x_n, x_{n+1})).$$

Since $k \in (0, 1)$ and $\varphi(t) < t$ for $t > 0$, we have $k\varphi(t) < kt < t$. Thus,

$$\mathfrak{d}(x_n, x_{n+1}) \leq k\varphi(\mathfrak{d}(x_n, x_{n+1})) < \mathfrak{d}(x_n, x_{n+1}),$$

which is a contradiction unless $\mathfrak{d}(x_n, x_{n+1}) = 0$. If $\mathfrak{d}(x_n, x_{n+1}) = 0$, then $x_n = x_{n+1}$, which implies x_n is a fixed point, contradicting our assumption that $x_n \neq x_{n+1}$. Therefore, this case implies $\mathfrak{d}(x_n, x_{n+1}) = 0$, meaning we found a fixed point.

Case B. Suppose $\mathfrak{M}(x_{n-1}, x_n) = \mathfrak{d}(x_{n-1}, x_n)$. Then

$$\mathfrak{d}(x_n, x_{n+1}) \leq k\varphi(\mathfrak{d}(x_{n-1}, x_n)).$$

Since $\varphi(t) < t$ for $t > 0$, we have

$$\mathfrak{d}(x_n, x_{n+1}) < k\mathfrak{d}(x_{n-1}, x_n).$$

Case C. Suppose $\mathfrak{M}(x_{n-1}, x_n) = \frac{\mathfrak{d}(x_{n-1}, x_{n+1})}{2s}$. Then

$$\mathfrak{d}(x_n, x_{n+1}) \leq k\varphi(\frac{\mathfrak{d}(x_{n-1}, x_{n+1})}{2s}).$$

Since $\varphi(t) < t$ for $t > 0$, we have

$$\mathfrak{d}(x_n, x_{n+1}) < k\frac{\mathfrak{d}(x_{n-1}, x_{n+1})}{2s}.$$

Using the b-triangle inequality, $\mathfrak{d}(x_{n-1}, x_{n+1}) \leq s[\mathfrak{d}(x_{n-1}, x_n) + \mathfrak{d}(x_n, x_{n+1})]$ we have

$$\mathfrak{d}(x_n, x_{n+1}) \quad < \quad \frac{k}{2s}s[\mathfrak{d}(x_{n-1}, x_n) + \mathfrak{d}(x_n, x_{n+1})] = \frac{k}{2}[\mathfrak{d}(x_{n-1}, x_n) + \mathfrak{d}(x_n, x_{n+1})].$$

Let $d_n = \mathfrak{d}(x_n, x_{n+1})$. Then $d_n < \frac{k}{2}(d_{n-1} + d_n)$. Hence,

$$d_n < \frac{k}{2-k}d_{n-1}.$$

Since $k \in (0, 1)$, we have $2 - k > 1$, so $\frac{k}{2-k} < k < 1$.

Let $\lambda = \max\{k, \frac{k}{2-k}\}$. Since $k \in (0, 1)$, we have $\frac{k}{2-k} < k$, and thus $\lambda = k$. Therefore, in all cases where $\mathfrak{d}(x_n, x_{n+1}) > 0$, we have:

$$\mathfrak{d}(x_n, x_{n+1}) \leq k\varphi(\mathfrak{M}(x_{n-1}, x_n)) < k\mathfrak{M}(x_{n-1}, x_n).$$

This implies $\mathfrak{d}(x_n, x_{n+1}) < k\mathfrak{d}(x_{n-1}, x_n)$ if $\mathfrak{M}(x_{n-1}, x_n) = \mathfrak{d}(x_{n-1}, x_n)$, or $\mathfrak{d}(x_n, x_{n+1}) < \frac{k}{2-k}\mathfrak{d}(x_{n-1}, x_n)$ if $\mathfrak{M}(x_{n-1}, x_n) = \frac{\mathfrak{d}(x_{n-1}, x_{n+1})}{2s}$. Therefore, we have $\mathfrak{d}(x_n, x_{n+1}) \leq \gamma\mathfrak{d}(x_{n-1}, x_n)$ for some $\gamma \in (0, 1)$. Thus,

$$\mathfrak{d}(x_n, x_{n+1}) \leq \lambda^n \mathfrak{d}(x_0, x_1)$$

where $\lambda = \max\{k, \frac{k}{2-k}\} < 1$. This implies $\lim_{n\to\infty} \mathfrak{d}(x_n, x_{n+1}) = 0$.

Now, we show that $\{x_n\}$ is a Cauchy sequence. For any $m > n$, using the b-triangle inequality repeatedly, we get

$$\begin{aligned}
\mathfrak{d}(x_n, x_m) &\leq s\mathfrak{d}(x_n, x_{n+1}) + s^2\mathfrak{d}(x_{n+1}, x_{n+2}) + \cdots + s^{m-n}\mathfrak{d}(x_{m-1}, x_m) \\
&= \sum_{j=n}^{m-1} s^{j-n+1}\mathfrak{d}(x_j, x_{j+1}) \\
&\leq \sum_{j=n}^{m-1} s^{j-n+1}\lambda^j\mathfrak{d}(x_0, x_1) = \mathfrak{d}(x_0, x_1)\sum_{j=n}^{m-1} s^{j-n+1}\lambda^j \\
&= \mathfrak{d}(x_0, x_1)\cdot s\sum_{j=n}^{m-1} s^{j-n}\lambda^j = \mathfrak{d}(x_0, x_1)\cdot s\lambda^n \sum_{j=0}^{m-n-1} (s\lambda)^j.
\end{aligned}$$

Since, $s\max(k, \frac{k}{2-k}) < 1$, then the series converges, and since $\lambda^n \to 0$ as $n \to \infty$, it follows that $\mathfrak{d}(x_n, x_m) \to 0$ as $n, m \to \infty$. Thus, $\{x_n\}$ is a Cauchy sequence. By the completeness of $\mathfrak{X}$, there exists $x \in \mathfrak{X}$ such that $\lim_{n\to\infty} x_n = x$. We now show that x is a fixed point of $\mathfrak{T}$. We have

$$\mathfrak{d}(\mathfrak{T}x, x) \leq s[\mathfrak{d}(\mathfrak{T}x, \mathfrak{T}x_n) + \mathfrak{d}(\mathfrak{T}x_n, x)] = s[\mathfrak{d}(\mathfrak{T}x, \mathfrak{T}x_n) + \mathfrak{d}(x_{n+1}, x)],$$

and

$$\mathfrak{d}(\mathfrak{T}x, \mathfrak{T}x_n) \leq k\varphi(\mathfrak{M}(x, x_n)),$$

where

$$\mathfrak{M}(x, x_n) = \max\Big\{\mathfrak{d}(x, x_n), \mathfrak{d}(x, \mathfrak{T}x), \mathfrak{d}(x_n, \mathfrak{T}x_n), \frac{d(x, \mathfrak{T}x_n) + \mathfrak{d}(x_n, \mathfrak{T}x)}{2s}\Big\}.$$

As $n \to \infty$, $\mathfrak{M}(x, x_n) \to \max\{0, \mathfrak{d}(x, \mathfrak{T}x), 0, \frac{0+0}{2s}\} = \mathfrak{d}(x, \mathfrak{T}x)$. Therefore,

$$\mathfrak{d}(\mathfrak{T}x, x) \leq s[k\varphi(\mathfrak{d}(x, \mathfrak{T}x)) + 0] = s\,k\,\varphi(\mathfrak{d}(x, \mathfrak{T}x)).$$

If $\mathfrak{d}(x, \mathfrak{T}x) > 0$, then,

$$\mathfrak{d}(\mathfrak{T}x, x) \leq s\,k\,\varphi(\mathfrak{d}(x, \mathfrak{T}x)) < s\,k\,\mathfrak{d}(x, \mathfrak{T}x) < \mathfrak{d}(x, \mathfrak{T}x).$$

which is a contradiction. This implies $\mathfrak{d}(x, \mathfrak{T}x) = 0$, so $\mathfrak{T}x = x$, and x is a fixed point of $\mathfrak{T}$.
To prove uniqueness, suppose x and y are two fixed points of $\mathfrak{T}$, i.e., $\mathfrak{T}x = x$ and $\mathfrak{T}y = y$. Then using (1),

$$\mathfrak{d}(x, y) = \mathfrak{d}(\mathfrak{T}x, \mathfrak{T}y) \leq k\varphi(\mathfrak{M}(x, y)),$$

where

$$\begin{aligned}
\mathfrak{M}(x, y) &= \max\Big\{\mathfrak{d}(x, y), \mathfrak{d}(x, \mathfrak{T}x), \mathfrak{d}(y, \mathfrak{T}y), \frac{\mathfrak{d}(x, \mathfrak{T}y) + \mathfrak{d}(y, \mathfrak{T}x)}{2s}\Big\} \\
&= \max\{\mathfrak{d}(x, y), 0, 0, \mathfrak{d}(x, y)/s\} = \mathfrak{d}(x, y).
\end{aligned}$$

Thus, $\mathfrak{d}(x, y) \leq k\varphi(\mathfrak{d}(x, y)) < k\mathfrak{d}(x, y)$. Since $k < 1$, this implies $\mathfrak{d}(x, y) = 0$, and therefore $x = y$. Hence, the fixed point is unique. This completes the proof. □

Example 3.4. Let $\mathfrak{X} = [0,1]$ and $\mathfrak{d}(x,y) = |x-y|^2$. Then $(\mathfrak{X}, \mathfrak{d})$ is a complete $\mathfrak{b}$-metric space with $s = 2$. Define $\mathfrak{T} : \mathfrak{X} \to \mathfrak{X}$ by $\mathfrak{T}x = x/4$. Let $\varphi(t) = t/2$ and $k = 1/2$. Then for $x, y \in \mathfrak{X}$,

$$\begin{aligned} \mathfrak{d}(\mathfrak{T}x, \mathfrak{T}y) &= |\mathfrak{T}x - \mathfrak{T}y|^2 = |x/4 - y/4|^2 = \frac{1}{16}|x-y|^2 \\ &\leq \frac{1}{2}\varphi(|x-y|^2) = k\varphi(\mathfrak{d}(x,y)). \end{aligned}$$

Since $\varphi(t) < t$ for all $t > 0$ and $k = 1/2 < 1$, $\mathfrak{T}$ is a φ-nonlinear contraction. By Theorem 3.3, $\mathfrak{T}$ has a unique fixed point in $\mathfrak{X}$, which is $x = 0$.

Remark 3.5. Theorem 3.2 generalizes the Banach contraction principle in $\mathfrak{b}$-metric spaces by considering a more general nonlinear contraction condition. The choice of the function φ allows for greater flexibility in the contraction requirement. Furthermore, it reduces to the standard Banach contraction principle when $\varphi(t) = t$ and $s = 1$.

4. Application to Differential Equations

In this section, we demonstrate the application of the fixed point theorems established in Section 3 to prove the existence and uniqueness of solutions to certain types of nonlinear differential equations.

Consider the following nonlinear differential equation:

$$\begin{cases} x'(t) = \mathfrak{f}(t, x(t)), \quad t \in \mathfrak{I} = [a,b], \\ x(a) = x_0, \end{cases} \tag{4.1}$$

where $\mathfrak{f} : \mathfrak{I} \times \mathbb{R} \to \mathbb{R}$ is a continuous function. We assume that $\mathfrak{f}$ satisfies the following Lipschitz condition:

$$|\mathfrak{f}(t,x) - \mathfrak{f}(t,y)| \leq \mathfrak{L}_t |x-y|,$$

for all $t \in \mathfrak{I}$ and $x, y \in \mathbb{R}$, where $\mathfrak{L}_t$ is a continuous function on $\mathfrak{I}$. We define an operator $\mathfrak{T} : \mathfrak{C}(\mathfrak{I}, \mathbb{R}) \to \mathfrak{C}(\mathfrak{I}, \mathbb{R})$ by

$$(\mathfrak{T}x)(t) = x_0 + \int_a^t \mathfrak{f}(\tau, x(\tau))d\tau, \tag{4.2}$$

where $\mathcal{C}(\mathfrak{I}, \mathbb{R})$, the space of continuous functions from $\mathfrak{I}$ to $\mathbb{R}$, equipped with the $\mathfrak{b}$-metric $d(x,y) = \sup_{t\in[a,b]} |x(t) - y(t)|^2$, with $s = 2$. It is clear that $x \in \mathfrak{C}([a,b], \mathbb{R})$ is a solution of the Eq. (2) if and only if x is a fixed point of $\mathfrak{T}$.

Theorem 4.1. Suppose that the Lipschitz constant $\mathfrak{L}_t$ satisfies

$$\max_{t\in\mathfrak{I}} \Big(\int_a^t \mathfrak{L}_\tau d\tau \Big)^2 < 1/4.$$

Then the Eq. (2) has a unique solution in $\mathfrak{C}(\mathfrak{I}, \mathbb{R})$.

Proof. We show that the operator $\mathfrak{T}$ defined by (3) is a φ-nonlinear contraction on $\mathcal{C}(\mathfrak{I}, \mathbb{R})$. For any $x, y \in \mathcal{C}(\mathfrak{I}, \mathbb{R})$, we have

$$\begin{aligned} |(\mathfrak{T}x)(t) - (\mathfrak{T}y)(t)| &= \Big| \int_a^t [\mathfrak{f}(\tau, x(\tau)) - \mathfrak{f}(\tau, y(\tau))]d\tau \Big| \leq \int_a^t |\mathfrak{f}(\tau, x(\tau)) - \mathfrak{f}(\tau, y(\tau))|d\tau \\ &\leq \int_a^t \mathfrak{L}_\tau |x(\tau) - y(\tau)|d\tau \leq \max_{\tau\in\mathfrak{I}} |x(\tau) - y(\tau)| \int_a^t L_\tau d\tau \end{aligned}$$

Therefore,

$$|(\mathfrak{T}x)(t)-(\mathfrak{T}y)(t)|^2 \leq \Big(\int_a^t \mathfrak{L}_\tau d\tau\Big)^2 (\max_{\tau\in\mathfrak{I}}|x(\tau)-y(\tau)|)^2 = \Big(\int_a^t \mathfrak{L}_\tau d\tau\Big)^2 \max_{\tau\in\mathfrak{I}}|x(\tau)-y(\tau)|^2.$$

Taking the maximum over $t\in[a,b]$, we get

$$\max_{t\in\mathfrak{I}}|(\mathfrak{T}x)(t)-(\mathfrak{T}y)(t)|^2 \leq \max_{t\in\mathfrak{I}}\Big(\int_a^t \mathfrak{L}_\tau d\tau\Big)^2 \max_{\tau\in\mathfrak{I}}|x(\tau)-y(\tau)|^2,$$

which implies

$$\mathfrak{d}(\mathfrak{T}x,\mathfrak{T}y) \leq \max_{t\in\mathfrak{I}}\Big(\int_a^t \mathfrak{L}_\tau \mathfrak{d}\tau\Big)^2 \mathfrak{d}(x,y).$$

Let $k = 4\max_{t\in\mathfrak{I}}\Big(\int_a^t \mathfrak{L}_\tau d\tau\Big)^2$ and choose $\varphi(t)=t/4$. Then $k<1$, $\varphi(t)<t$ for all $t>0$ and $\varphi(0)=0$. Then we have $\mathfrak{d}(\mathfrak{T}x,\mathfrak{T}y)\leq k\varphi(\mathfrak{d}(x,y))$. Hence, $\mathfrak{T}$ is a φ-nonlinear contraction with $\mathfrak{M}(x,y)=\mathfrak{d}(x,y)$. Therefore, by Theorem 3.2, $\mathfrak{T}$ has a unique fixed point in $\mathfrak{C}(\mathfrak{I},\mathbb{R})$, which is the unique solution to the Eq. (2). This completes the proof. □

Example 4.2. Consider the differential equation $x'(t)=t\,x(t)$, $x(0)=1$, for $t\in[0,1]$. Here, $\mathfrak{f}(t,x)=t\,x$. Then $|\mathfrak{f}(t,x)-\mathfrak{f}(t,y)|=|t\,x-t\,y|=t|x-y|$. So $\mathfrak{L}_t=t$. Then $\int_0^t \mathfrak{L}_\tau d\tau=\int_0^t \tau d\tau = t^2/2$. Thus, $\max_{t\in[0,1]}(\int_0^t \mathfrak{L}_\tau d\tau)^2 = \max_{t\in[0,1]}(t^2/2)^2 = 1/4$. By Theorem 4.1, the differential equation has a unique solution in $\mathfrak{C}([0,1],\mathbb{R})$.

References

[1] Bakhtin I. A., The contraction mapping principle in quasimetric spaces. Func. An., Gos. Ped. Inst. Unianowsk, (1989) Vol. 30, P. 26–37.

[2] Czerwik S., Contraction mappings in $\mathfrak{b}$-metric spaces. Acta Mathematica et Informatica Universitatis Ostraviensis, (1993) Vol. 1, P. 5–11.

[3] Czerwik S., Nonlinear set-valued contraction mappings in $\mathfrak{b}$-metric spaces. Atti Seminario Matematico e Fisico dell'Universita di Modena, (1998) Vol. 46, P. 263–276.

[4] Roshan J. R., Hussain N., Ozturk V., Radenović S., Fixed point theorems for weakly contractive mappings in $\mathfrak{b}$-metric spaces. Applied Mathematics and Computation, (2016) Vol. 275, P. 275–285.

[5] Gurro A., Jleli M., Fixed-point results on $\mathfrak{b}$-metric spaces and applications to integral equations. Journal of Fixed Point Theory and Applications, (2019) Vol. 21, Iss. 2, P. 1–13.

[6] Barman D., Sarkar K., Tiwary K., Fixed point theorems in complex valued $\mathfrak{b}$-metric spaces. Math. Moravica, (2023) Vol. 27, Iss. 1, P. 85–96.

[7] Berinde V., Păcurar M., The early developments in fixed point theory on b-metric spaces: a brief survey and some important related aspects. Carpathian Journal of Mathematics, (2022) Vol. 38, Iss. 3, P. 523–538.

[8] Devi D., Debnath P., Fixed points of two interpolative cyclic contractions in $\mathfrak{b}$-metric spaces. Heliyon, (2025) Vol. 11, P. e41667.

[9] Dung N. V., On Cirić type theorems in $\mathfrak{b}$-metric spaces. Publ. Inst. Math. (2023) Vol. 113, Iss. 127, P. 99–108.

[10] Komachalia A. H. A., Gabaa Y. U., Aphanea M., Yildirim I., Refinement on fixed point results in metric and $\mathfrak{b}$-metric spaces. Filomat, (2023) Vol. 37, Iss. 22, P. 7581–7588.

[11] Navascués M. A., Mohapatra R. N., Fixed point dynamics in a new type of contraction in $\mathfrak{b}$-metric spaces. Symmetry, (2024) Vol. 16, P. 506.

[12] Huang H., Qian X., Common fixed point of nonlinear contractive mappings. AIMS Mathematics, (2022) Vol. 8, Iss. 1, P. 607–621.

[13] Pant R., Fixed point theorems for nonlinear contractions with applications to iterated function systems. Appl. Gen. Topol. (2018) Vol. 19, Iss. 1, P. 163–172.

[14] Sintunavarat W., Nonlinear integral equations with new admissibility types in $\mathfrak{b}$-metric spaces. J. Fix. Point. Theory. Appli. (2016) Vol. 18, Iss. 2, P. 397–416.

[15] Boriceanu M., Fixed point theory for multivalued generalized contraction on a set with two $\mathfrak{b}$-metric. Studia Universitatis Babes-Bolyai Mathematica, (2009) Vol. 54, P. 3–14.

Mahdi H.S.,
Department of Mathematics, Faculty of Basic Sciences,
Babol Noshirnani University of Technology, Babol, Iran
e-mail: s.h.rasouli@nit.ac.ir

Rasouli S.H.,
Department of Mathematics, Faculty of Basic Sciences,
Babol Noshirnani University of Technology, Babol, Iran
e-mail: s.h.rasouli@nit.ac.ir

Mohammadzadeh B.,
Department of Mathematics, Faculty of Basic Sciences,
Babol Noshirnani University of Technology, Babol, Iran
e-mail: s.h.rasouli@nit.ac.ir

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 159–166
DOI: 10.29229/uzmj.2026-2-19

Mixed piecewise constant argument and the Runge-Kutta method for numerically solving first-order differential equations

Muminov M., Jumaev Z.

Abstract. This work introduces an efficient methodology for approximating solutions to first-order non-linear differential equations. The approach is based on formulating a differential equation with piecewise constant arguments associated with the parameters of the classical Runge-Kutta (RK) discrete equations depended on a positive integer parameter n. It is demonstrated that the constructed equation has a unique piecewise-smooth solution and for sufficiently large values of n, this solution approximating solution to the original problem. Numerical results are provided through examples, demonstrating the efficiency and high accuracy of the proposed method.

Keywords: Initial value problem, Differential equation with piecewise constant argument, Approximated solution, Runge-Kutta method, Absolute error
MSC (2020): 34A12, 34A34, 65L05, 65L06

1. Introduction

The finite element approach has been widely applied to nonlinear equations [1]. Both mesh-based and meshless methods have advanced the solving of linear and nonlinear ODEs, PDEs, and fractional differential equations [2], [3], [4].

A finite-time stability for hybrid dynamical systems with deviating argument is explored using a hybrid control scheme in [5]. The reliability of the Non-Fixed Step-Size Algorithm offering efficient integration over varying intervals for stiff differential systems is confirmed through validation of its properties [6], [7], [8]. Finding exact periodic solutions of first-order non-homogeneous differential equations with piecewise constant arguments, providing conditions and explicit solutions are given in [9], [10].

Recent published papers has demonstrated numerical methods for solving initial value problems using various approaches, including Euler's method, the Taylor series method, the Adomian decomposition method, and the Runge-Kutta method [11]. In addition, conventional Taylor and fourth-order Runge-Kutta (RK4) methods have been compared in [12].

In [13], the approximation of initial value problems was addressed through the introduction of an innovative and enhanced method inspired by Taylor's approach, which outperforms the Runge-Kutta (RK) method in terms of accuracy.

In this paper, we propose an effective method for constructing an approximate solution to a class of nonlinear ordinary differential equations of the form

$$y'(t) = f(t, y(t)), \qquad y(a) = y_0, \qquad t \in [a, b], \tag{1.1}$$

where $f(\cdot, \cdot)$ is a real-valued function continuous in $t \in [a, b]$ and Lipschitz continuous in $y \in \mathbb{R}$ (with Lipschitz constant independent from t), and y_0 is a given real constant. The approach involves constructing the initial value problem as a differential equation with piecewise constant arguments by using parameters of the seventh-order RK discrete equation, parametrized by a positive integer n. It is established that the resulting equation admits a unique piecewise-smooth solution, which serves as an approximation to the original problem when n becomes sufficiently large.

To demonstrate the practical effectiveness of the proposed method, several numerical experiments are conducted. The results are compared with the results of [11], [12] and [13] showing that the suggested approach yields high accuracy and exhibits rapid convergence.

2. Differential equations with piecewise constant arguments

Let n be a positive integer, and define the partition points as $t_k = kh + a$, where $h = \frac{b-a}{n}$ and $k = 0, 1, 2, \ldots, n$.

This study focuses on the constructing of differential equation with piecewise constant arguments using the coefficients of a seventh-order Runge-Kutta method comprising thirteen stages for solving initial value problems [14]. The main idea is based on the following sketch: the numbers $K(t_k)$ in (2.1) are constructed so that the function $F_n(t)$, defined in (2.4) converges uniformly to $f(t, y(t))$ on $[a, b]$ as $n \to \infty$, where $y(t)$ is a regular solution of (1.1).

The numerical approximation of the solution to (1.1) using the classical seventh-order Runge–Kutta method is expressed as [15], [14]:

$$\frac{y_k - y_{k-1}}{h} = K(t_k), \tag{2.1}$$

where h denotes the step size, and $K(t_k)$ represents the Runge-Kutta increment function evaluated at t_k, where

$$K(t_k) = \frac{31}{720}k_1(t_k) + \frac{16}{75}k_6(t_k) + \frac{16807}{79200}k_7(t_k)$$
$$+\frac{16807}{79200}k_8(t_k) + \frac{243}{1760}k_9(t_k) + \frac{243}{1760}k_{12}(t_k) + \frac{31}{720}k_{13}(t_k), \tag{2.2}$$

$k_1(t_k) = f(t_k, y(t_k))$,

$k_2(t_k) = f(t_k + \frac{1}{4}h, y(t_k) + h(\frac{1}{4}k_1(t_k)))$,

$k_3(t_k) = f(t_k + \frac{1}{12}h, y(t_k) + h(\frac{5}{72}k_1(t_k) + \frac{1}{72}k_2(t_k)))$,

$k_4(t_k) = f(t_k + \frac{1}{8}h, y(t_k) + h(\frac{1}{32}k_1(t_k) + \frac{3}{32}k_3(t_k)))$,

$k_5(t_k) = f(t_k + \frac{2}{5}h, y(t_k) + h(\frac{106}{125}k_1(t_k) - \frac{408}{125}k_3(t_k) + \frac{352}{125}k_4(t_k)))$,

$k_6(t_k) = f(t_k + \frac{1}{2}h, y(t_k) + h(\frac{1}{48}k_1(t_k) + \frac{8}{33}k_4(t_k) + \frac{125}{528}k_5(t_k)))$,

$k_7(t_k) = f(t_k + \frac{6}{7}h, y(t_k) + h(-\frac{1263}{2401}k_1(t_k) + \frac{39936}{26411}k_4(t_k) - \frac{64125}{26411}k_5(t_k) + \frac{5520}{2401}k_6(t_k)))$,

$k_8(t_k) = f(t_k + \frac{1}{7}h, y(t_k) + h(\frac{37}{392}k_1(t_k) + \frac{1625}{9408}k_5(t_k) - \frac{2}{15}k_6(t_k) + \frac{61}{6720}k_7(t_k)))$,

$k_9(t_k) = f(t_k + \frac{2}{3}h, y(t_k) + h(\frac{17176}{25515}k_1(t_k) - \frac{47104}{25515}k_4(t_k) + \frac{1325}{504}k_5(t_k) - \frac{41792}{25515}k_6(t_k) + \frac{20237}{145800}k_7(t_k) + \frac{4312}{6075}k_8(t_k)))$,

$k_{10}(t_k) = f(t_k + \frac{2}{7}h, y(t_k) + h(-\frac{23834}{180075}k_1(t_k) - \frac{77824}{1980825}k_4(t_k) - \frac{636635}{633864}k_5(t_k) + \frac{254048}{300125}k_6(t_k) - \frac{183}{7000}k_7(t_k) + \frac{8}{11}k_8(t_k) - \frac{324}{3773}k_9(t_k)))$,

$k_{11}(t_k) = f(t_k + h, y(t_k) + h(\frac{12733}{7600}k_1(t_k) - \frac{20032}{5225}k_4(t_k) + \frac{456485}{80256}k_5(t_k) - \frac{42599}{7125}k_6(t_k) + \frac{339227}{912000}k_7(t_k) - \frac{1029}{4180}k_8(t_k) + \frac{1701}{1408}k_9(t_k) + \frac{5145}{2432}k_{10}(t_k)))$,

$k_{12}(t_k) = f(t_k + \frac{1}{3}h, y(t_k) + h(-\frac{27061}{204120}k_1(t_k) + \frac{40448}{280665}k_4(t_k) - \frac{1353775}{1197504}k_5(t_k) + \frac{17662}{25515}k_6(t_k) - \frac{71687}{1166400}k_7(t_k) + \frac{98}{225}k_8(t_k) + \frac{1}{16}k_9(t_k) + \frac{3773}{11664}k_{10}(t_k)))$,

$k_{13}(t_k) = f(t_k + h, y(t_k) + h(\frac{11203}{8680}k_1(t_k) - \frac{38144}{11935}k_4(t_k) + \frac{2354425}{458304}k_5(t_k) - \frac{84046}{16275}k_6(t_k) + \frac{673309}{1636800}k_7(t_k) + \frac{4704}{8525}k_8(t_k) + \frac{9477}{10912}k_9(t_k) - \frac{1029}{992}k_{10}(t_k) + \frac{729}{341}k_{12}(t_k)))$.

For each positive integer n, we construct a differential equation incorporating a piecewise constant argument, which serves as an approximation to the nonlinear differential equation given in (1.1), and is associated with the corresponding initial value problem.

$$y'(t) = F_n(t), \quad y(t_0) = y_0, \quad t \in (a, b), \tag{2.3}$$

where

$$F_n(t) = K(t_{k-1}) \quad \text{for} \quad t \in [t_{k-1}, t_k), \quad k = 1, \ldots, n, \tag{2.4}$$

$$y_k = y(t_k) = \lim_{t \to t_k - 0} \int_{t_{k-1}}^{t} F_n(s) ds = K(t_{k-1}) \frac{1}{n},$$

By the construction of $K(t_{k-1})$, we have

$$K(t_{k-1}) - f(t_{k-1}, y_{k-1}) \to 0 \quad \text{as} \quad n \to \infty. \tag{2.5}$$

A solution to equation (2.3) is characterized in the following manner, as described in [16], [17].

Definition 2.1. A function $y(t) := y_n(t)$ is said to be a solution of the initial value problem (2.3) if the following conditions are fulfilled:

(i) $y(t)$ is continuous on the interval $[a, b]$;

(ii) The derivative $y'(t)$ exists and is continuous on $[a, b]$, except possibly at the points $t = t_k$, where $k = 0, 1, \ldots, n$, at which the one-sided derivatives exist;

(iii) The function $y(t)$ satisfies the differential equation in (2.3) for all $t \in (a, b)$, possibly excluding the points $t = t_k$, where $k = 0, 1, \ldots, n$.

Theorem 2.2. *For every positive integer n, the initial value problem* (2.3) *admits a unique solution $y(t) := y_n(t)$, as defined in* (3.5).

Let $y(t)$ be a solution of the initial value problem defined in (1.1). Then there exists $d > 0$ such that it satisfies the boundedness condition

$$|y(t)| \leq d.$$

Moreover, we assume that the function $f(t, y)$ has compact support $D = [a, b] \times [-2d, 2d]$, that is,

$$f(t, y) = 0 \quad \text{for} \quad (t, y) \notin D. \tag{2.6}$$

We note that under the given assumptions, the function $f(t, y)$ is bounded. Consequently, the corresponding solutions $y_n(t)$ are uniformly bounded as well. That is, there exists a positive constant C such that

$$|y_n(t)| \leq C. \tag{2.7}$$

The next theorem establishes that the function $y_n(t)$ serves as an approximate solution to the initial value problem (1.1).

Theorem 2.3. *For any given $\varepsilon > 0$, there exists a positive integer $n_0 = n(\varepsilon)$ such that for all integers $n > n_0$, the following inequality holds:*

$$\sup_{t \in [a,b]} |y_n'(t) - f(t, y_n(t))| < \varepsilon, \tag{2.8}$$

where $y_n(t)$ denotes the solution of the initial value problem (2.3).

3. Proof of the Main results

Proof of Theorem 2.2. Let n be a fixed positive integer. For $t \in [t_0, t_1)$, equation (2.3) takes the form:

$$y'(t) = K(t_0). \tag{3.1}$$

Integrating both sides of (3.1) yields:

$$y(t) = \int_{t_0}^{t} K(t_0)\, ds + y(t_0), \qquad t \in [t_0, t_1). \tag{3.2}$$

Since the function $y(t)$ is continuous on the interval $[t_0, t_1)$, the left-hand limit at $t = t_1$ exists

$$y(t_1) = \lim_{t \to t_1 - 0} y(t). \tag{3.3}$$

Now, let $y(t)$ denote the solution to equation (2.3) on the interval $[t_{k-2}, t_{k-1})$ with the initial condition $y(t_{k-2}) = \lim\limits_{t \to t_{k-2} - 0} y(t)$.

Let

$$y(t_{k-1}) = \lim_{t \to t_{k-1} - 0} y(t). \tag{3.4}$$

Further, using the initial condition (3.4), we integrate equation (2.3) over the interval $[t_{k-1}, t_k)$:

$$y(t) = K(t_{k-1})(t - t_{k-1}) + y(t_{k-1}), \qquad t \in [t_{k-1}, t_k), \tag{3.5}$$

for $k = 2, \ldots, n$.

By construction, the function $y(t)$ defined by (3.5) is differentiable on each subinterval

$$(a, t_1) \cup (t_1, t_2) \cup \cdots \cup (t_{n-1}, b),$$

and continuous on $[a, b]$. Hence, by construction, the function $y(t)$ is the unique solution to equation (2.3).

Proof of Theorem 2.3. Let y_n be the solution of equation (2.3). Then, for $t \in [t_k, t_{k+1})$, the following identity holds:

$$y_n'(t) - f(t, y_n(t)) = K(t_{k-1}) - f(t, y_n(t)).$$

We estimate the right-hand side as follows:

$$|K(t_{k-1}) - f(t, y_n(t))| \le |K(t_{k-1}) - f(t_{k-1}, y_n(t_{k-1}))| + |f(t_{k-1}, y_n(t_{k-1})) - f(t, y_n(t))|.$$

According to equation (2.5), there exists a constant $C > 0$ such that

$$|K(t_{k-1}) - f(t_{k-1}, y_n(t_{k-1}))| \le \frac{C}{n}$$

for sufficiently large n.

By equation (3.5), the solution $y_n(t)$ on $[t_k, t_{k+1})$ can be written as:

$$y_n(t) = K(t_{k-1})(t - t_{k-1}) + y(t_{k-1}),$$

so that

$$f\big(t, y_n(t)\big) = f\big(t, K(t_{k-1})(t - t_{k-1}) + y(t_{k-1})\big), \quad t \in [t_k, t_{k+1}).$$

Since the function f is differentiable on the domain D, there exists a constant $F > 0$ such that

$$|f(t, y_n(t_{k-1})) - f(t, y_n(t))| \le \frac{F}{n}, \tag{3.6}$$

for all $t \in [t_{k-1}, t_k)$ and sufficiently large n.

Combining the above estimates yields:

$$\big|y_n'(t) - f(t, y_n(t))\big| \le \frac{C + F}{n}, \tag{3.7}$$

for all $t \in [t_{k-1}, t_k)$, $k = 1, \ldots, n$, and sufficiently large n.

This completes the proof of the theorem.

4. Algorithmic Summary

The following steps outline the computational procedure for approximating the solution of the initial value problem:

Input: Prescribed initial condition.

Step 1: Compute the value of $K(t_0)$ using equation (2.2);

Step 2: Determine the function $y_n(t)$ over the interval $t \in [t_0, t_1)$ via equation (3.2);

Step 3: Evaluate the one-sided limit $y_n(t_1) = \lim_{t \to t_1 - 0} y_n(t)$;

Step 4: For each subinterval, calculate $K(t_{k-1})$ using equation (2.2);

Step 5: Apply formula (3.5) to compute $y_n(t)$ on the interval $t \in [t_{k-1}, t_k)$ for $k = 2, \ldots, n$, and evaluate the left-hand limit $y_n(t_{k-1}) = \lim_{t \to t_{k-1} - 0} y_n(t)$;

Step 6: Construct the approximate solution $y_n(t)$ using the piecewise-defined formula (3.5);

Step 7: Evaluate the maximum absolute error between the exact solution $y(t_{k-1})$ and the numerical approximation $y_n(t_{k-1})$ at the mesh points:

$$\text{MaxErr} = \max_{1 \le k \le n} \left| y(t_{k-1}) - y_n(t_{k-1}) \right|;$$

Output: If the computed maximum error is within the prescribed tolerance, the procedure terminates successfully.

5. Numerical Results

To obtain an approximate solution of the initial value problem (1.1) based on Theorem 2.3, it is essential to compute the function y_n, which is explicitly defined by formula (3.5).

Example 1. Consider the initial value problem described in [13]:

$$y'(t) = te^{y(t)}, \quad 0 \le t \le 0.7, \quad y(0) = 1.$$

The parameters are $N = 10$, $h = 0.07$, and $t_i = 0.07i$. The exact solution corresponding to this problem is given by

$$y(t) = -\ln\left(e^{-1} - \frac{1}{2}t^2\right).$$

The absolute error results by using a hybrid technique combining the piecewise constant argument approach with the Runge-Kutta method are given in Table 1. The comparison with the numerical results obtained via the classical Runge-Kutta (RK) method, the Obreschkoff method, and our proposed approach highlights the accuracy and efficiency of the proposed scheme.

Table 1. Comparison of absolute errors in the Runge-Kutta (RK) method of order 6 [13], the Obreschkoff method (OM) of order 8 [13] and present method applied in Example 1 with step size $h = 0.07$.

t_i	**RK Error [13]**	**OM Error [13]**	**Present Method**
0.0	0.0000000	0.0000000	0.0000000
0.07	7.9991×10^{-10}	1.9000×10^{-13}	8.2486×10^{-16}
0.14	1.8513×10^{-9}	9.5000×10^{-13}	4.1093×10^{-15}
0.21	8.5394×10^{-9}	2.9800×10^{-12}	1.1712×10^{-14}
0.28	2.0817×10^{-8}	8.5900×10^{-12}	2.8994×10^{-14}
0.35	4.1761×10^{-8}	2.5740×10^{-11}	7.1300×10^{-14}
0.42	7.7709×10^{-8}	8.6570×10^{-11}	1.8795×10^{-13}
0.49	1.3675×10^{-7}	3.5061×10^{-10}	5.6835×10^{-13}
0.56	2.3260×10^{-7}	1.8740×10^{-9}	2.1447×10^{-12}
0.64	3.1537×10^{-7}	1.5364×10^{-8}	1.1346×10^{-11}
0.70	7.3768×10^{-7}	2.6095×10^{-7}	8.8373×10^{-11}

Example 2. Consider the Abel-type nonlinear differential equation [11]

$$y'(t) = t + 3ty + 3ty^2 + ty^3,$$

subject to the initial condition $y(0) = 0$. The exact solution to this problem is expressed as

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

TABLE 2. The table of absolute errors at the point t_i obtained by Euler, Taylor, Adomian, RK th4 [11] and Present methods for Example 2.

t	Exact	Abs. Err Euler [11]	Abs. Err Taylor [11]	Abs. Err Adomian [11]	Abs. Err RK 4th [11]	Abs. Err Present method
0.1	5.0378e-03	1.0000e+00	7.5063e-03	2.7352e-07	6.1705e-06	7.9062e-15
0.2	2.0621e-02	5.1505e-01	1.9136e-02	1.7543e-05	6.4452e-06	5.2515e-14
0.3	4.8285e-02	3.6614e-01	3.1291e-02	2.0138e-04	7.2163e-06	1.5983e-13
0.4	9.1089e-02	3.0348e-01	4.5789e-02	1.1539e-03	8.6288e-06	4.1544e-13
0.5	1.5470e-01	2.7892e-01	6.4632e-02	4.5820e-03	1.0966e-05	1.2183e-12
0.6	2.5000e-01	2.7911e-01	9.1203e-02	1.4693e-02	1.4457e-05	6.574e-12
0.7	4.0028e-01	3.0334e-01	1.3228e-01	4.1529e-02	1.8686e-05	1.0107e-10
0.8	6.6667e-01	3.6210e-01	2.0385e-01	1.0966e-01	1.8517e-05	4.3633e-09
0.9	1.2942e+00	4.9242e-01	3.5428e-01	2.8525e-01	1.0507e-03	9.3657e-07

Example 3. The Generalized Taylor method is employed to approximate the solution of the initial value problem [12]:

$$y'(t) = e^{t-y(t)}, \quad 0 \le t \le 1, \quad y(0) = 1,$$

where the parameters are specified as $N = 10$, $h = 0.1$, and $t_i = 0.1i$. The exact analytical solution for this problem is given by

$$y(t) = \ln(e^t + e^{-1} - 1).$$

Moreover, the performance of the proposed approximation is compared against several classical numerical methods to assess its accuracy and efficiency.

TABLE 3. Comparison of absolute errors for the Modified Euler (MEM), Trapezoid (TM), Midpoint (MM), Taylor and RK methods of order 4, alongside the Generalized Taylor method (GTM) [12] and Present method for Example 3, with a step size of $h = 0.1$.

t_i	MEM [12]	TM [12]	MM [12]	Taylor [12]	RKM [12]	GTM [12]	Present method
0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.1	2.8×10^{-5}	2.8×10^{-5}	8.7×10^{-6}	9.0×10^{-9}	4.4×10^{-9}	5.2×10^{-10}	2.9×10^{-12}
0.2	5.6×10^{-5}	5.6×10^{-5}	1.8×10^{-5}	1.6×10^{-8}	9.2×10^{-9}	9.5×10^{-10}	3.1×10^{-12}
0.3	8.3×10^{-5}	8.3×10^{-5}	2.9×10^{-5}	2.2×10^{-8}	1.3×10^{-8}	1.3×10^{-9}	2.9×10^{-12}
0.4	1.1×10^{-4}	1.1×10^{-4}	4.1×10^{-5}	2.6×10^{-8}	2.0×10^{-8}	1.4×10^{-9}	2.5×10^{-12}
0.5	1.4×10^{-4}	1.4×10^{-4}	5.3×10^{-5}	2.7×10^{-8}	2.5×10^{-8}	1.5×10^{-9}	2.2×10^{-12}
0.6	1.6×10^{-4}	1.6×10^{-4}	6.7×10^{-5}	2.6×10^{-8}	3.1×10^{-8}	1.4×10^{-9}	1.9×10^{-12}
0.7	1.9×10^{-4}	1.9×10^{-4}	8.0×10^{-5}	2.3×10^{-8}	3.8×10^{-8}	1.2×10^{-9}	1.6×10^{-12}
0.8	2.1×10^{-4}	2.1×10^{-4}	9.4×10^{-5}	1.8×10^{-8}	4.4×10^{-8}	8.8×10^{-10}	1.4×10^{-12}
0.9	2.3×10^{-4}	2.3×10^{-4}	1.1×10^{-4}	1.2×10^{-8}	5.0×10^{-8}	4.7×10^{-10}	1.2×10^{-12}
1.0	2.5×10^{-4}	2.5×10^{-4}	1.2×10^{-4}	4.2×10^{-9}	5.6×10^{-8}	1.9×10^{-11}	1.1×10^{-12}

It is evident that the present method yields substantially higher accuracy in comparison to the other numerical schemes under consideration. The comparative analysis of the approximate solutions obtained via the present method and alternative approaches, along with their corresponding absolute errors, is presented in Table 3.

Example 4. Let

$$y'(t) = e^{y^2(t)} + \cos(t) - e^{\sin^2(t)}, \quad 0 \le t \le 1, \quad y(0) = 0.$$

The exact solution of this IVP is

$$y(t) = \sin(t).$$

The results for $N = 10$, $h = 0.1$, and $t_k = 0.1k$ using the proposed method are presented in Table 4.

TABLE 4. The absolute errors in the Runge-Kutta (RK) method of order 7 in Example 4 with step size $h = 0.1$.

t_i	**Exact**	**Approximate**	**Absolute error**
0.0	0.00000000000000	0.00000000000000	0.0000000
0.1	0.0998334166468281	0.0998334166468285	3.5143×10^{-16}
0.2	0.1986693307950612	0.1986693307950617	5.1645×10^{-16}
0.3	0.2955202066613396	0.2955202066613394	1.8762×10^{-16}
0.4	0.3894183423086505	0.3894183423086475	3.0056×10^{-15}
0.5	0.4794255386042030	0.4794255386041950	7.9623×10^{-15}
0.6	0.5646424733950354	0.5646424733950275	7.8677×10^{-15}
0.7	0.6442176872376911	0.6442176872377125	2.1443×10^{-14}
0.8	0.7173560908995228	0.7173560908996445	1.2174×10^{-13}
0.9	0.7833269096274834	0.7833269096278118	3.2845×10^{-13}
1.0	0.8414709848078965	0.8414709848086330	7.3648×10^{-13}

6. Conclusion

A novel numerical scheme is introduced for approximating solutions of differential equations. The proposed methodology is based on constructing an associated differential equation incorporating a piecewise constant argument, using the parameters seventh-order RK method, which corresponds to the original problem. A solution to such ordinary differential equations with piecewise constant arguments is defined as a piecewise smooth function, depended on a positive integer n. It is demonstrated that, for sufficiently large values of n, the solution of the constructed equation closely approximates the solution of the original initial value problem.

In Examples 1-4, using the proposed method, the maximum absolute errors between the exact and the approximate solutions at the points t_k, are demonstrated, with the results of the recently published papers, in Tables 1-4, respectively. The comparisons show the accuracy and efficiency of the proposed technique.

References

[1] Nagel J. R., Introduction to the Finite Element Method. (2012).

[2] Ahsan M., Meshless analysis of nonlocal boundary value problems in anisotropic and inhomogeneous media. Mathematics, (2020) Vol. 8, Iss. 11, P. 2045.

[3] Fu Z.-J., A robust kernel-based solver for variable-order time fractional PDEs under 2D/3D irregular domains. Applied Mathematics Letters, (2019) Vol. 94, P. 105–111.

[4] Fu Z., Localized collocation schemes and their applications. Acta Mechanica Sinica, (2022) Vol. 38, Iss. 7, P. 422167.

[5] Xi Q., Liu X., Finite-time stability and controller design for a class of hybrid dynamical systems with deviating argument. Nonlinear Analysis: Hybrid Systems, (2021) Vol. 39, P. 100952.

[6] Sunday J., A pair of three-step hybrid block methods for the solutions of linear and nonlinear first-order systems. European Journal of Mathematics and Statistics, (2022) Vol. 3, Iss. 1, P. 14–25.

[7] Sunday J., Numerical integration of stiff differential systems using non-fixed step-size strategy. Symmetry, (2022) Vol. 14, Iss. 8, P. 1575.

[8] Sunday J., Shokri A., Marian D., Variable step hybrid block method for the approximation of Kepler problem. Fractal and Fractional, (2022) Vol. 6, Iss. 6, P. 343.

[9] Muminov M. I., Jumaev Z. Z., Exact periodic solutions of second-order differential equations with piecewise constant arguments. Advances in Mathematics: Scientific Journal, (2021) Vol. 10, Iss. 9, P. 3113–3128.

[10] Muminov M. I., Radjabov T. A., Existence conditions for 2-periodic solutions to a non-homogeneous differential equations with piecewise constant argument. Examples and Counterexamples, (2024) Vol. 5, P. 100145.

[11] Almajbri T. F. A., Mohammed A. B., Numerical Studies for Solving Abel's Differential Equation. Malaysian Journal of Industrial Technology, (2024) Vol. 8, Iss. 4, P. 93–104.

[12] Alomari M. W., Euler-Maclaurin method for approximating solutions of initial value problems. International Journal of Robotics and Control Systems, (2025) Vol. 5, Iss. 1, P. 366–380.

[13] Alomari M. W., Revolutionizing Numerical Approximations: A Novel Higher-Order Implicit Method vs. Runge-Kutta for Initial Value Problems. Journal of Robotics and Control (JRC), (2025) Vol. 6, Iss. 2, P. 624–637.

[14] Verner J. H., Explicit Runge–Kutta methods with estimates of the local truncation error. SIAM Journal on Numerical Analysis, (1978) Vol. 15, Iss. 4, P. 772–790.

[15] Trikkaliotis G. D., Gousidou-Koutita M. C., Production of the reduction formula of seventh order runge-kutta method with step size control of an ordinary differential equation. Applied Mathematics, (2022) Vol. 13, Iss. 4, P. 325–337.

[16] Muminov M. I., Jumaev Z. Z., Piecewise Constant Argument Method for Nonlinear Systems of Differential Equations. UiTM International Conference on Mathematical Sciences, Singapore, (2024).

[17] Muminov M. I., Jumaev Z. Z., On the Approximation Algorithms for Solving Generalized Bernoulli Differential Equations Using Piecewise Constant Argument Method. International Journal of Applied and Computational Mathematics, (2025) Vol. 11, Iss. 2, P. 45.

Mukhiddin I. Muminov,
Faculty of Mathematics, Samarkand State University, Samarkand, Uzbekistan,
V.I Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences, Tashkent, Uzbekistan
e-mail: mimuminov@gmail.com

Zafar Z. Jumaev,
Faculty of Intelligent Systems and Computer Technologies, Samarkand State University, Samarkand, Uzbekistan
e-mail: zafarlangar8708@gmail.com

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 167–173
DOI: 10.29229/uzmj.2026-2-20

Geometry of orbits of Hamiltonian vector fields

Narmanov A., Parmonov H.

Abstract. Hamilton systems play a fundamental role in mechanics and mathematics, which usually generated by Hamiltonian vector fields.The geometry and topology of vector field orbits is not only an important branch of modern geometry, but also has important applications in many fields of science and engineering. This article studies the geometric characteristics of the leaves of the foliation generated by the orbits of Hamiltonian vector fields.

Keywords: Hamiltonian vector field, Gauss curvature, Gauss torsion, foliation
MSC (2020): 37J05, 53C12

1. Introduction

Let M be a smooth Riemannian manifold of dimension n with the Riemannian metric g, ∇ is the Levi-Civita connection, $< \cdot, \cdot >$ is inner product defined by the Riemannian metric g.

We denote by $V(M)$ the set of all smooth vector fields defined on M, through $[X, Y]$ Lie bracket of vector fields $X, Y \in V(M)$. The set $V(M)$ is a Lie algebra with Lie bracket.

Let's consider some set $D \subset V(M)$, which contains finite or infinite number of smooth vector fields. For a point $x \in M$ through $t \to X^t(x)$ we will denote the integral curve of a vector field X passing through a point x at $t = 0$. Map $t \to X^t(x)$ is defined in some domain $I(x) \subset R$, which generally depends on field X and point x.

Definition 1.1. The orbit $L(x)$ of a system D of vector fields through a point x is the set of points y in M such that there exist $t_1, t_2, ..., t_k \epsilon R$ and vector fields $X_1, X_2, ..., X_k \epsilon D$ such that

$$y = X_k^{t_k}(X_{k-1}^{t_k-1}(...(X_1^{t_1}(x)))),$$

where k is an arbitrary positive integer.

There exist many papers devoted to topology and geometry of orbits [1],[2],[3],[4].

In this paper we study the geometry of orbits of Hamiltonian vector fields.

Let us recall a notion of Hamiltonian vector field. In order to define notion of Hamiltonian vector field we need to define Poisson bracket of functions.

Definition 1.2. A *Poisson bracket* on a smooth manifold M is an operation that assigns a smooth real-valued function $\{F, H\}$ on M to each pair F, H of smooth, real-valued functions, with the basic properties:

(a)*Bilineality:*

$$\{\lambda F + \mu P, H\} = \lambda\{F, H\} + \mu\{P, H\},$$

$$\{F, \lambda H + \mu P\} = \lambda\{F, H\} + \mu\{F, P\},\ \lambda, \mu \in \mathbb{R};$$

(b)*Skew-Symmetry:*

$$\{F, H\} = -\{H, F\};$$

(c)*Jacobi Identity:*

$$\{\{F, H\}, P\} + \{\{P, F\}, H\} + \{\{H, P\}, F\} = 0;$$

(c)*Leibniz' Rule:*

$$\{F, H \cdot P\} = \{F, H\} \cdot P + H \cdot \{F, P\}.$$

(Here $\cdot$ denotes the ordinary multiplication of real-valued functions.)

A manifold M with a Poisson bracket is called a Poisson manifold, the bracket defining a Poisson structure on M The notion of a Poisson manifold is slightly more general than that of a simplistic manifold, or manifold with Hamiltonian structure; in particular, the underlying manifold M need not be even-dimensional. This is borne out by the standard examples from classical mechanics.

Let M be the Euclidean space R^m, with coordinates

$$(p, q, z) = (p^1, ..., p^n, q^1, ..., q^n, z^1, ..., z^l),$$

where $m = 2n + l$.

If $F(p, q, z)$ and $H(p, q, z)$ are smooth functions, we define their Poisson bracket to be the function:

$$\{F, H\} = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial q^i} \frac{\partial H}{\partial p^i} - \frac{\partial F}{\partial p^i} \frac{\partial H}{\partial q^i} \right).$$

This bracket is clearly bilinear and skew-symmetric; the verifications of the Jacobi identity and the Leibniz rule are simple exercises. We note the particular bracket identities:

$$\{p^i, p^j\} = 0,\ \{q^i, q^j\} = 0,\ \{p^i, q^j\} = \delta^i_j,$$

$$\{p^i, z^k\} = \{q^i, z^k\} = \{z^t, z^k\} = 0.$$

in which i and j run from 1 to n, when t and k run from 1 to l. Here δ^i_j is the Kronecker symbol, which is 1 if $i = j$ and 0 otherwise.

Definition 1.3. Let M be a Poisson manifold and $H : M \to R$ a smooth function. The *Hamiltonian vector field* associated with H is the unique smooth vector field $sgradH$ on M satisfying

$$sgradH(F) = \{F, H\} = -\{H, F\}$$

for every smooth function $F : M \to R$.

The equations governing the flow of $sgradH$ are referred to as *Hamilton's equations* for the *Hamiltonian function* H.

In the case of the canonical Poisson bracket on R^m $(m = 2n + l)$, the Hamiltonian vector field to any $H(p, q, z)$, as clearly, corresponds

$$sgradH = \sum_{i=1}^{n} \left(\frac{\partial H}{\partial p^i} \frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i} \frac{\partial}{\partial p^i} \right)$$

The corresponding flow is obtained by integrating the system of ordinary differential equations

$$\frac{dp^i}{dt} = -\frac{\partial H}{\partial q^i},\ \ \frac{dq^i}{dt} = \frac{\partial H}{\partial p^i},\ \ i = 1, ..., n,$$

$$\frac{dz^l}{dt} = 0,\ \ l = 1, ..., l.$$

There is a fundamental connection between the Poisson bracket of two functions and the Lie bracket of their associated *Hamiltonian vector fields*, which forms the basis of much of the theory of Hamiltonian systems. It is well known following theorem [5]:

Theorem 1.4. *Let M be a Poisson manifold and $\{F, H\} : M \to R$ are smooth functions with corresponding Hamiltonian vector fields sgradF, sgradH. The Hamiltonian vector field associated with the Poisson bracket of F and H is, up to sign, the Lie bracket of the two Hamiltonian vector fields:*

$$sgrad\{F, H\} = [sgradH, sgardF].$$

The Hamiltonian vector field associated with $H(x)$ has the form

$$sgradH = \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\{x^i, x^j\}\frac{\partial H}{\partial x^j}\frac{\partial}{\partial x^i}\right).$$

Let $F(x)$ be a second smooth function. We obtain the basic formula

$$\{F, H\} = \sum_{i=1}^{n}\sum_{j=1}^{n}\{x^i, x^j\}\frac{\partial F}{\partial x^i}\frac{\partial H}{\partial x^j}$$

for the Poisson bracket.

This basic brackets $A^{ij}(x) = \{x^i, x^j\}$ $i, j = 1, ..., m$ are called the *structure functions* of the Poisson manifold M with respect to the given local coordinates.

A skew-symmetric $m \times m$ matrix $A(x) = \{A^{ij}(x)\}$ called the *structure matrix* of M. If the Poisson bracket is non-degenerate ($det(A^{ij}) \neq 0$ everywhere on M), then the Poisson manifold is called a *symplectic manifold*. The symplectic structure in this case has the form $\omega = A_{ij}dx^i \wedge dx^j$ where A_{ij} are components of the matrix inverse to (A^{ij}).

2. Main part

I.Let us consider functions $H_1, H_2 : \mathbb{R}^4 \to \mathbb{R}$ on the Euclidean four dimensional space $\mathbb{R}^4$ with cartezian coordinates p_1, p_2, q_1, q_2 which are given by the formulas

$$H_1 = \frac{1}{2}(p_1^2 + p_2^2 + q_1^2 + q_2^2)$$

$$H_2 = \frac{1}{2}(p_1^2 + p_2^2 - q_1^2 + q_2^2).$$

The Hamiltonian vector fields corresponding to H_1 and H_2 have following forms:

$$sgradH_1 = -q_1\frac{\partial}{\partial p^1} - q_2\frac{\partial}{\partial p^2} + p_1\frac{\partial}{\partial q^1} + p_2\frac{\partial}{\partial q^2}, sgradH_2 = q_1\frac{\partial}{\partial p^1} - q_2\frac{\partial}{\partial p^2} + p_1\frac{\partial}{\partial q^1} + p_2\frac{\partial}{\partial q^2}. \quad (2.1)$$

Theorem 2.1. *The orbits of vector fields (2.1) generate singular foliation regular leaf of which is three dimensional surface with zero Gauss curvature.*

Proof. We need invariant function from conditions

$$sgradH_1(F) = 0, sgradH_2(F) = 0 \quad (2.2)$$

and we get the function $F(p_1, p_2, q_1, q_2) = p_2^2 + q_2^2$ is a invariant function. For the Puasson bracket of functions H_1 va H_2 we have

$$\{H_1, H_2\} = 2p_1q_1.$$

From theorem 2.1 it follows that

$$[sgradH_2, sgradH_1] = sgrad\{H_1, H_2\} = -2p_1\frac{\partial}{\partial p^1} + 2q_1\frac{\partial}{\partial q^1}$$

If we denote by H_3 the function $\{H_1, H_2\}$ then we have

$$sgradH_3 = [sgradH_2, sgradH_1].$$

And also we have $sgradH_3(F) = 0$. It follows from [6],[5] vector fields $sgradH_1, sgradH_2, sgradH_3$ are tangent vector fields for every orbit of the system D. Let denote by $A(D)$ minimal Lie subalgebra of the algebra $V(R^4)$ containing the set D, and define subspace

$$A_x(D) = \{X(x) : X \in A(D)\}, x = (p_1, p_2, q_1, q_2) \in R^4\}$$

of tangent space at the point $x = (p_1, p_2, q_1, q_2)$. It is known, that

$$dimA_x(D) = \{X(x) : X \in A(D)\} \geq dimL(x)$$

where $L(x)$ is a orbit containing x [7]. If $p_1^2 + q_1^2 > 0$ and $p_2^2 + q_2^2 > 0$ then

$$dimA_x(D) = dimL(x) = 3.$$

It follows from here in this case the orbit $L(x)$ is three dimensional surface.

Let us consider connected component Φ of the surface defined by equation $p_2^2 + q_2^2 = c$. Then if $x \in \Phi$ the orbit $L(x)$ is a open subset of the surface Φ. Since the different orbits do not intersect and the set Φ is connected, we obtain that $L(x) = \Phi$.

Let us Φ parameterize the by following equations

$$\begin{cases} p_1 = u \\ p_2 = \sqrt{c}\cos s \\ q_1 = v \\ q_2 = \sqrt{c}\sin s \end{cases}$$

or by vector equation

$$r = r(u, v, s) = \{u,\ \sqrt{c}\cos s,\ v,\ \sqrt{c}\sin s\}.$$

We find matrixs of first and second quatric forms:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & c \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\sqrt{c} \end{pmatrix}.$$

In order to find principal curvatures we have following equation $det|B - \lambda A| = 0$. We have following values of principal curvatures

$$\lambda_1 = 0,\ \lambda_2 = 0,\ \lambda_3 = -\frac{1}{\sqrt{c}}$$

and Gauss curvature $\lambda_1 \cdot \lambda_2 \cdot \lambda_3 = 0$.

Following sets are singular leaves:

$L_1 = \{0, 0, 0, 0\}-$ zero dimensional orbit;

$L_2 = \{(0, p_2, 0, q_2) : p_2^2 + q_2^2 = c,\ c > 0\}-$ one dimensional orbit.

We also have following leaf: $L_3 = \{(p_1, 0, q_1, 0) : p_1^2 + q_1^2 > 0\}-$ two dimensional plane (p_1, q_1) without the point $(0, 0)$.

II. Now let us consider functions

$$H_1 = p_1q_1 + p_2q_2$$

$$H_2 = p_1q_2 + p_2q_1$$

and hamiltonian vector fields

$$sgradH_1 = -p_1\frac{\partial}{\partial p^1} - p_2\frac{\partial}{\partial p^2} + q_1\frac{\partial}{\partial q^1} + q_2\frac{\partial}{\partial q^2}, sgradH_2 = -p_2\frac{\partial}{\partial p^1} - p_1\frac{\partial}{\partial p^2} + q_2\frac{\partial}{\partial q^1} + q_1\frac{\partial}{\partial q^2}. \quad (2.3)$$

Theorem 2.2. *The orbits of vector fields (2.3) generate singular foliation regular leaf of which is two dimensional surface zero Gauss curvature and zero Gauss torsion.*

Proof. From the equality $\{H_1, H_2\} = 0$ we get

$$[sgradH_2, sgradH_1] = 0.$$

From the following conditions

$$sgradH_1(F_1) = 0,\ sgradH_2(F_1) = 0,\ sgradH_1(F_2) = 0,\ sgradH_2(F_2) = 0$$

we find following invariant functions

$$F_1(p_1, p_2, q_1, q_2) = p_1q_2 + p_2q_1, F_2(p_1, p_2, q_1, q_2) = p_1q_1 + p_2q_2.$$

Let us consider connected component Φ of the surface defined by equation

$$\begin{cases} p_1q_2 + p_2q_1 = c_1, c_1 \neq 0 \\ p_1q_1 + p_2q_2 = c_2, c_2 \neq 0 \end{cases}$$

If $x \in \Phi$ then

$$dimA_x(D) = dimL(x) = 2.$$

It follows from here in this case the orbit $L(x)$ is two dimensional surface. Since the surface Φ is also two-dimensional, the orbit $L(x)$ is an open subset of the surface Φ. Since the different orbits do not intersect and the set Φ is connected, we obtain that $L(x) = \Phi$.

We parameterize this surface by following equations

$$\begin{cases} p_1 = u \\ p_2 = v \\ q_1 = \dfrac{c_2v - c_1u}{v^2 - u^2} \\ q_2 = \dfrac{c_1v - c_2u}{v^2 - u^2} \end{cases} \tag{2.4}$$

or in vector form

$$r = r(u, v) = \{u,\ v,\ \frac{c_2v - c_1u}{v^2 - u^2},\ \frac{c_1v - c_2u}{v^2 - u^2}\}.$$

Using this equations we find coefficients of first quadric form

$$g_{11} = 1 + \frac{(c_1^2 + c_2^2)(u^4 + 6u^2v^2 + v^4) - 8c_1c_2uv(u^2 + v^2)}{(v^2 - u^2)^4}$$

$$g_{12} = g_{21} = \frac{c_1c_2(u^2 + 6u^2v^2 + v^4) - 2(c_1^2 + c_2^2)uv(u^2 + v^2)}{(v^2 - u^2)^4}$$

$$g_{22} = 1 + \frac{(c_1^2 + c_2^2)(u^4 + 6u^2v^2 + v^4) - 8c_1c_2uv(u^2 + v^2)}{(v^2 - u^2)^4},$$

we use denotation $c_1c_2 = \tilde{c}$, $c_1^2 + c_2^2 = \tilde{c}^2$ and also we find normal vectors

$$n_1 = \left\{ \frac{-(c_1 + c_2)}{\sqrt{2(c_1 + c_2)^2 + 2(u + v)^4}}, \frac{-(c_1 + c_2)}{\sqrt{2(c_1 + c_2)^2 + 2(u + v)^4}}, \right.$$

$$\left. \frac{(u + v)^2}{\sqrt{2(c_1 + c_2)^2 + 2(u + v)^4}}, \frac{(u + v)^2}{\sqrt{2(c_1 + c_2)^2 + 2(u + v)^4}} \right\}$$

$$n_2 = \left\{ \frac{c_1 - c_2}{\sqrt{2(c_1 - c_2)^2 + 2(v - u)^4}}, \frac{c_2 - c_1}{\sqrt{2(c_1 - c_2)^2 + 2(v - u)^4}}, \right.$$

$$\left.\frac{(v-u)^2}{\sqrt{2(c_1-c_2)^2+2(v-u)^4}}, \frac{-(v-u)^2}{\sqrt{2(c_1-c_2)+2(v-u)^4}}\right\}.$$

Now we find two second quadric forms

$$II_1 = b_{ij}du^i du^j, II_2 = c_{ij}du^i du^j,$$

where

$$b_{11} = b_{12} = b_{22} = \frac{\sqrt{2}(c_1+c_2)}{(u+v)\sqrt{(c_1+c_2)^2+(u+v)^4}}$$

and

$$c_{11} = -c_{12} = c_{22} = \frac{\sqrt{2}(c_2-c_1)}{(v-u)\sqrt{(c_1-c_2)^2+(v-u)^4}}$$

Let us recall some notions on the geometry of two dimensional submanifolds of four dimensional Euclidean space [8],[9],[10], [11],[12].

Through a point $x_0 \in F^n$ in some tangent direction τ we draw a curve γ lying on F^n. Let s be the length of the arc on γ. Consider the curvature vector k of the curve γ: $k = r_{ss}$. *The normal curvature vector* k_N of a surface F_n in the direction τ at a point x_0 is called the projection of the curvature vector k of a curve γ onto the normal space N_{x_0}.

Let

$$u^i = u^i(s),$$

$i = 1, \ldots, n$,– be the equation of γ. Then

$$r_{ss} = r_{u^i u^j}\frac{du^i}{ds}\frac{du^j}{ds} + r_{u^i}\frac{d^2u^i}{ds^2}\cdot.$$

The projection of the vector k onto the normal space N_{x_0} has the form

$$k_N = \sum_{\sigma=1}^{p}(kn_\sigma)n_\sigma = \sum_{\sigma=1}^{p}(r_{ss}n_\sigma)n_\sigma.$$

Using the expression for r_{ss}, we can write

$$k_N = \sum_{\sigma=1}^{p}(r_{u^i u^j}n_\sigma)\frac{du^i}{ds}\frac{du^j}{ds}n_\sigma = \sum_{\sigma=1}^{p}\frac{II^\sigma n_\sigma}{I}.$$

It follows from here normal curvature vector does not depend from curve γ, it depends only from directionτ.

The *indicatrix of the normal curvature* of a submanifold at a point x_0 is the set of points N_{x_0} of the endpoints of the normal curvature vectors $k_N(\tau)$, taken for each tangent direction τ and plotted from the point x_0.

The indicatrix of the normal curvature is an ellips. Let us denote by a, b the half axis of the *indicatrix of the normal curvature*. We define the Gaussian torsion χ_G of a two-dimensional surface in four-dimensional space as

$$\sigma_G = 2ab.$$

The *Gaussian torsion* of two dimensional surface F in $\mathbb{R}^4$ is a function that assigns a value to each point of the surface, calculated by the following formula

$$\sigma_G = \frac{\langle n_{1u}, n_{2v}\rangle - \langle n_{2u}, n_{1v}\rangle}{\sqrt{g_{11}g_{22}-g_{12}^2}}$$

In our case, the normal vectors n_1, n_2 are non- constant vectors.

Let $F^2 \subset \mathbb{R}^4$ be a 2-dimensional smooth surface embedded in 4-dimensional Euclidean space. At each point $p \in F$, the tangent space T_pF is 2-dimensional, and the normal space N_pF is also 2-dimensional.

Let $\{n_1, n_2\}$ be an orthonormal basis of the normal space N_pF. For each normal vector n_i, we can define: the second fundamental form h^{n_i}.

Then the **Gaussian curvature** $K(p)$ of the surface F at the point p is given by:

$$K(p) = \det(h^{\nu_1}) + \det(h^{\nu_2})$$

where: h^{n_i} is the matrix of the second fundamental form in the direction of n_i. It follows Gaussian curvature of two dimensional surface F is given by the formula

$$K(p) = \frac{b_{11}b_{22} - b_{12}^2}{g_{11}g_{22} - g_{12}^2} + \frac{c_{11}c_{22} - c_{12}^2}{g_{11}g_{22} - g_{12}^2}.$$

Calculations with formulas gives that

$$K(p) = 0, \sigma_G = 0.$$

Remark: Unlike the case in $\mathbb{R}^3$, where the Gaussian curvature is an intrinsic invariant depending only on the metric of the surface, in $\mathbb{R}^4$ the Gaussian curvature also depends on the extrinsic geometry that is, on how the surface is embedded into the ambient space.

–

References

[1] Azamov A., Narmanov A., On the Limit Sets of Orbits of Systems of Vector Fields. Differential Equations, (2004) Vol. 40, Iss. 2, P. 257–260.

[2] Narmanov A., Zoyidov A. N., On the group of diffeomorphisms of foliated manifolds. Vestnik Udmurtskogo Universiteta. Matematika, Mekhanika, Komp'yuternye Nauki, (2020) Vol. 30, Iss. 1, P. 49–58.

[3] Narmanov A. Y., Stability of completely controllable systems. Differential Equations, (2000) Vol. 36, Iss. 10, P. 1475–1483.

[4] Narmanov A., The transversal structure of controllability sets for symmetric control systems. Differential Equations, (1996) Vol. 32, Iss. 6, P. 786–789.

[5] Olver P., Applications of Lie Groups to Differential Equations. Springer, (1993).

[6] Narmanov O., Invariant solutions of the two-dimensional heat equation. Bulletin of Udmurt University. Mathematics, Mechanics, Computer Science, (2019) Vol. 29, Iss. 1, P. 52–60.

[7] Sussman H., Orbits of families of vector fields and integrability of distributions. Transactions of the AMS, (1973) Vol. 180, P. 171–188.

[8] Aminov Y. A., Szajewska M. G., Gaussian torsion of a 2-dimensional surface defined implicitly in 4-dimensional Euclidean space. Sbornik: Mathematics, (2004) Vol. 195, Iss. 11, P. 1545–1556.

[9] Dorodnitsyn A., Knyazeva I., Svirshchevskii S., Group properties of the heat equation with source in the two-dimensional and three-dimensional cases. Differential Equations, (1983) Vol. 19, Iss. 7, P. 1215–1223.

[10] Fomenko V., Classification of Two-Dimensional Surfaces with Zero Normal Torsion in Four-Dimensional Spaces of Constant Curvature. Mathematical Notes, (2004) Vol. 75, Iss. 5, P. 690–701.

[11] Fomenko V., Some properties of two-dimensional surfaces with zero normal torsion in E^4. Sbornik: Mathematics, (1979) Vol. 35, Iss. 2, P. 251–265.

[12] Kadomcev S., A study of certain properties of normal torsion of a two-dimensional surface in four-dimensional space. Akad. Nauk SSSR Vsesojuz. Inst. Nauch. i Tehn. Informacii, Moscow, (1975).

Narmanov A.Y.,
Department of Geometry, National University of Uzbekistan, Tashkent,Uzbekistan,
e-mail: narmanov@yandex.ru

Parmonov H.F.,
Department of Differential Equations, Bukhara State University of Uzbekistan, Bukhara,Uzbekistan
e-mail: hparmonov@mail.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 174–186
DOI: 10.29229/uzmj.2026-2-21

Anti-associative dendriform algebras

Normatov Z.

Abstract. The general operadic approach to splitting algebraic operations was developed in [1]. By splitting the product in a given algebraic variety $\mathcal{C}$, notion of $\mathcal{C}$-dendriform algebras was systematically studied in [2]. This article aims to study "anti-associative dendriform algebras", which offer an approach to addressing anti-associativity. These algebras are defined by two operations whose sum is anti-associative. Furthermore, the notion of $\mathcal{O}$-operators on anti-associative algebras is presented as a tool to interpret anti-associative dendriform algebras. Moreover, anti-associative algebras with nondegenerate Connes cocycles admit compatible anti-associative dendriform algebra structures.

Keywords: Lie algebra, pre-Lie algebra, dendriform algebra, Jacobi-Jordan algebra, pre-Jacobi-Jordan algebra, $\mathcal{O}$-operator
MSC (2020): 17A01, 17C50

1. Motivation

Lie algebras were introduced by Sophus Lie in the 1870s to formalize the study of infinitesimal transformations. Since then, they have become fundamental tools in a wide range of areas, including representation theory, differential geometry, algebraic geometry, number theory, quantum mechanics, and gauge theory.

A vector space A with a bilinear operation $*$ is called a Lie admissible algebra if $(A, [-,-])$ is a Lie algebra, where $[-,-]$ is the commutator. In particular, an *associative algebra* $(A, \cdot)$ which is a vector space A with a bilinear operation $\cdot$ satisfiying $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ is a Lie admissible algebra. Hence, there exist a functor from the category **As** of associative algebras to the category **Lie** of Lie algebras, i.e.

$$\mathbf{As} \to \mathbf{Lie}.$$

Furthermore, pre-Lie algebras, also known by various names such as left-symmetric algebras, quasi-associative algebras, or Vinberg algebras, are also Lie-admissible algebras. These algebras were first introduced by A. Cayley in 1896 as a type of rooted tree algebra. Hence, one can define a functor from the category **Pre-Lie** of all pre-Lie algebras to the category **Lie** of all Lie algebras. Thus, one has the following diagram

$$\begin{array}{ccc} & & \mathbf{Pre\text{-}Lie} \\ & & \downarrow \\ \mathbf{As} & \to & \mathbf{Lie} \end{array} \tag{1.1}$$

Question. Is it possible to extend the diagram (1.1) to obtain a commutative diagram?

To give an affirmative answer one should put the category **Dend** of all dendriform algebras on the left-top side (see [3, 4, 5]) i.e.

$$\begin{array}{ccc} \mathbf{Dend} & \longrightarrow & \mathbf{Pre\text{-}Lie} \\ \downarrow & & \downarrow \\ \mathbf{As} & \longrightarrow & \mathbf{Lie} \end{array}$$

The notion of dendriform algebras was introduced by Loday [6] in 1995 with the motivation from algebraic K-theory and has been studied quite extensively with connections to several areas in mathematics and physics, including operads, homology, Hopf algebras, Lie and Leibniz algebras, combinatorics, arithmetic and quantum field theory and so on (see [7] and the references therein).

Recall that a *dendriform algebra* $(A, \prec, \succ)$ is a vector space A with two binary operations denoted by $\prec$ and $\succ$ satisfying (for any $x, y, z \in A$)

$$(x \prec y) \prec z = x \prec (y \cdot z), \ \ (x \succ y) \prec z = x \succ (y \prec z), \ \ x \succ (y \succ z) = (x \cdot y) \succ z,$$

where $x \cdot y = x \prec y + x \succ y$. Note that $(A, \cdot)$ is an associative algebra as a direct consequence and the notion of dendriform algebras can be introduced as an approach of "splitting the associativity".

The central idea of the paper [2] was to generalize dendriform algebras by requiring that the operations $\prec$ and $\succ$ induce on $A \times A$ an algebra structure lying in a chosen category $\mathcal{C}$. The authors introduced the notion of $\mathcal{C}$-dendriform algebras which can be applied to any variety $\mathcal{C}$. This provides a unified framework in which usual dendriform algebras of of J.L-Loday appear as a special case when $\mathcal{C}$ is the category of associative algebras.

The goal of this paper is to study $\mathcal{C}$-dendriform algebras in the case $\mathcal{C}$ is the category of anti-associative algebras i.e. anti-associative dendriform algebras. This characterization parallels that of dendriform algebras.

An anti-associative algebra $(A, *)$ is a nonassociative algebra whose multiplication satisfies the identity

$$a * (b * c) + (a * b) * c = 0, \quad \forall a, b, c \in A.$$

On the one hand, if we consider the commutator, then the resulting algebra will not be a Lie algebra. But considering the anti-commutator $[a, b] = a*b+b*a$ derives Jacobi-Jordan algebras (see [8]), which seem to be similar to Lie algebras at first sight, but these algebras are quite different [9]. This class of algebras appeared in the literature under different names from different viewpoints, for example "Lie-Jordan algebras" [10], "mock-Lie algebras" [11]. Thus, one has the functor from the category **AAs** of anti-associative algebras to the category **J-J** of Jacobi-Jordan algebras i.e.

$$\mathbf{AAs} \to \mathbf{J\text{-}J}$$

On the other hand, similar to the construction of pre-Lie algebras, pre-Jacobi-Jordan algebras are constructed in [12] as a left-anti-symmetric algebra with respect to anti-associator. Hence, we have

$$\begin{array}{ccc} & & \mathbf{Pre\text{-}J\text{-}J} \\ & & \downarrow \\ \mathbf{AAs} & \to & \mathbf{J\text{-}J} \end{array} \tag{1.2}$$

It is then natural to ask whether the diagram (1.2) can be extended to a commutative diagram. In this paper we give an affirmative answer to this question and the category **AAs-Dend** of all **anti-associative dendriform algebras** play a left-top category in the commutative diagram i.e.

$$\begin{array}{ccc} \mathbf{AAs\text{-}Dend} & \dashrightarrow & \mathbf{Pre\text{-}J\text{-}J} \\ \vdots & & \downarrow \\ \mathbf{AAs} & \longrightarrow & \mathbf{J\text{-}J} \end{array} \tag{1.3}$$

The classification of any class of algebras is a fundamental but highly challenging problem. It is often one of the first issues encountered when attempting to understand the structure of a particular class of algebras. Specifically, the classification of algebras of dimension n (up to isomorphism) within a variety defined by a given family of polynomial identities is a classic question in the theory of non-associative algebras. Numerous studies have addressed the classification of low-dimensional algebras across various varieties of both associative and non-associative algebras [13, 14, 15, 16, 17, 18, 19, 20].

The paper is organized as follows. In Section 2, we recall the notion of anti-associative dendriform algebras as an approach of splitting the anti-associativity and prove that the diagram (1.3) is commutative. The notions of $\mathcal{O}$-operators and anti-Rota–Baxter operators on anti-associative algebras are introduced to interpret anti-associative dendriform algebras in Section 3. In Section 4, the relationships between anti-associative dendriform algebras and Connes cocycles on anti-associative algebras are given. In the last Section 5, we introduce various notions of nilpotency for anti-associative dendriform algebras and demonstrate that these definitions are equivalent. Moreover, we give the classification of 2-dimensional anti-associative dendriform algebras over the field of complex numbers.

2. anti-associative dendriform algebras

In this section, we prove that the diagram (1.3) is commutative. Moreover, using a bimodule structure on a vector space, we define an anti-associative algebra structure on the direct sum of the vector space with itself by using the operations of an anti-associative dendriform algebra.

Definition 2.1 ([9])**.** An algebra $(A, \diamond)$ over a field $\mathbb{F}$ is called a Jacobi–Jordan algebra if it satisfies the following two identities

$$x \diamond y - y \diamond x = 0,$$
$$x \diamond (y \diamond z) + y \diamond (z \diamond x) + z \diamond (x \diamond y) = 0,$$

for all $x, y, z \in A$.

In other words, Jacobi–Jordan algebras are commutative algebras which satisfy the Jacobi identity. In the case of Lie algebras, the product is not commutative but anticommutative.

Recall that $(A, \circ)$ is called a **Jacobi-Jordan-admissible algebra** [12, 8], where A is a vector space with a bilinear operation $\circ : A \otimes A \to A$, if the bilinear operation $[-,-] : A \otimes A \to A$ defined by

$$[x, y] = x \circ y + y \circ x, \ \ \forall x, y \in A,$$

makes $(A, [-,-])$ a Jacobi-Jordan algebra. In this case, $(A, [-,-])$ is called the **sub-adjacent Jacobi-Jordan algebra** of $(A, \circ)$ and denoted by $(\mathfrak{J}(A), [-,-])$. Obviously, an anti-associative algebra is a Jacobi-Jordan-admissible algebra.

Another example of Jacobi-Jordan-admissible algebras is pre-Jacobi-Jordan algebras.

Definition 2.2. [12] A pre-Jacobi-Jordan algebra is a vector space A together with a bilinear multiplication $\cdot : A \otimes A \to A$, such that for any $x, y, z \in A$, the anti-associator $AAs(x, y, z) := (x \cdot y) \cdot z + x \cdot (y \cdot z)$ is skew-symmetric in x, y, i.e. $AAs(x, y, z) = -AAs(y, x, z)$ or equivalently

$$(x \cdot y) \cdot z + x \cdot (y \cdot z) = -(y \cdot x) \cdot z - y \cdot (x \cdot z).$$

We now give a definition analogous to that of an associative admissible algebra [21].

Definition 2.3. Let A be a vector space with two bilinear operations

$$\succ : A \otimes A \to A, \quad \prec : A \otimes A \to A.$$

Define a bilinear operation $*$ as

$$x * y = x \succ y + x \prec y, \quad \forall x, y \in A. \tag{2.1}$$

The triple $(A, \succ, \prec)$ is called an anti-associative admissible algebra if $(A, *)$ is an anti-associative algebra. In this case, $(A, *)$ is called the associated anti-associative algebra of $(A, \succ, \prec)$.

Remark 2.4. The triple $(A, \succ, \prec)$ is an anti-associative admissible algebra if and only if the following equation holds:

$$\begin{aligned} &(x \succ y) \succ z + (x \prec y) \succ z + (x \succ y) \prec z + (x \prec y) \prec z \\ &= -x \succ (y \succ z) - x \succ (y \prec z) - x \prec (y \succ z) - x \prec (y \prec z), \quad \forall x, y, z \in A. \end{aligned} \tag{2.2}$$

Definition 2.5. Let A be a vector space with two bilinear operations $\succ$ and $\prec$. The triple $(A, \succ, \prec)$ is called a **anti-associative dendriform algebra** if the following equations hold:

$$(x * y) \succ z = -x \succ (y \succ z), \tag{2.3}$$
$$x \prec (y * z) = -(x \prec y) \prec z, \tag{2.4}$$
$$x \succ (y \prec z) = -(x \succ y) \prec z, \quad \forall x, y, z \in A, \tag{2.5}$$

where the bilinear operation $*$ is defined by Eq. (2.1).

Example 2.6. Let $(A, \prec, \succ)$ be a one-dimensional anti-associative dendriform algebra with basis element e. Assume that $e \succ e = \alpha e$ and $e \prec e = \beta e$, where $\alpha, \beta \in \mathbb{F}$. Then, by equations (2.3) – (2.5), we obtain $\alpha(2\alpha + \beta) = 0$ and $\beta(2\beta + \alpha) = 0$. Hence $\alpha = \beta = 0$, and every one- dimensional anti-associative dendriform algebra is trivial.

Theorem 2.7. *Let $(A, \succ, \prec)$ be an anti-associative dendriform algebra. Then the followings hold.*

(i). Define a bilinear operation $*$ by Eq. (2.1)*. Then $(A, *)$ is an anti-associative algebra, called the* **associated anti-associative algebra** *of $(A, \succ, \prec)$. Furthermore, $(A, \succ, \prec)$ is called a* **compatible anti-associative dendriform algebra structure** *on $(A, *)$.*

(ii). The bilinear operation $\circ : A \otimes A \to A$ given by

$$x \circ y = x \succ y + y \prec x, \quad \forall x, y \in A,$$

defines a pre-Jacobi-Jordan algebra, called the **associated pre-Jacobi-Jordan algebra** *of $(A, \succ, \prec)$.*

(iii). Both $(A, *)$ and $(A, \circ)$ have the same sub-adjacent Jacobi-Jordan algebra $(\mathfrak{g}(A), [-,-])$ defined by

$$[x, y] = x \succ y + x \prec y + y \succ x + y \prec x, \quad \forall x, y \in A.$$

Moreover, the commutative diagram (2) holds.

Proof. (i) is straightforward.

(ii) Let $x, y, z \in A$. Then we have

$$\begin{aligned}(x \circ y) \circ z =& (x \succ y + y \prec x) \succ z + z \prec (x \succ y + y \prec x)\\ =& (x \succ y) \succ z + (y \prec x) \succ z + z \prec (x \succ y) + z \prec (y \prec x),\\ x \circ (y \circ z) =& x \succ (y \succ z + z \prec y) + (y \succ z + z \prec y) \prec x\\ =& x \succ (y \succ z) + x \succ (z \prec y) + (y \succ z) \prec x + (z \prec y) \prec x.\end{aligned}$$

By swapping x and y, we have

$$\begin{aligned}(y \circ x) \circ z &= (y \succ x) \succ z + (x \prec y) \succ z + z \prec (y \succ x) + z \prec (x \prec y),\\ y \circ (x \circ z) &= y \succ (x \succ z) + y \succ (z \prec x) + (x \succ z) \prec y + (z \prec x) \prec y.\end{aligned}$$

Using Eqs. (2.3)–(2.5), we obtain

$$\begin{aligned}x \circ (y \circ z) + y \circ (x \circ z) =& x \succ (y \succ z) + x \succ (z \prec y) + (y \succ z) \prec x + (z \prec y) \prec x\\ &+ y \succ (x \succ z) + y \succ (z \prec x) + (x \succ z) \prec y + (z \prec x) \prec y\\ =& x \succ (y \succ z) + (z \prec y) \prec x + y \succ (x \succ z) + (z \prec x) \prec y\\ =& -(x \succ y + x \prec y) \succ z - z \prec (y \succ x + y \prec x)\\ &-(y \succ x + y \prec x) \succ z - z \prec (x \succ y + x \prec y)\\ =& -(y \circ x) \circ z - (x \circ y) \circ z.\end{aligned}$$

Moreover, we have

$$x \circ y + y \circ x = x \succ y + y \prec x + x \prec y + y \succ x = x * y + y * x, \quad \forall x, y \in A.$$

Thus $(A, \circ)$ is a Jacobi-Jordan-admissible algebra and hence a pre-Jacobi-Jordan algebra.

(iii). It is straightforward. Note that it also appears in the proof of (ii). □

Let $(A, *)$ be an anti-associative algebra. Recall that a **bimodule** of $(A, *)$ is a triple (V, l, r) consisting of a vector space V and linear maps $l, r : A \to \mathrm{End}_{\mathbb{F}}(\mathrm{V})$ such that

$$l(x * y)v = -l(x)(l(y)v), \quad r(x * y)v = -r(y)(r(x)v), \quad l(x)(r(y)v) = -r(y)(l(x)v), \quad \forall x, y \in A, v \in V.$$

In particular, (A, L_*, R_*) is a bimodule of $(A, *)$, where $L_*, R_* : A \to \mathrm{End}_{\mathbb{F}}(A)$ are two linear maps defined by $L_*(x)(y) = R_*(y)(x) = x * y$ for all $x, y \in A$ respectively.

Suppose that $(A,*)$ is an anti-associative algebra. Let V be a vector space and $l,r:A\to \mathrm{End}_{\mathbb{F}}(V)$ be linear maps. Then (V,l,r) is a bimodule of $(A,*)$ if and only if there is an anti-associative algebra structure on the direct sum $A\oplus V$ of vector spaces with the following bilinear operation, still denoted by $*$:

$$(x,u)*(y,v)=(x*y,l(x)v+r(y)u),\quad \forall x,y\in A,u,v\in V.$$

We denote this anti-associative algebra by $A\ltimes_{l,r}V$.

Theorem 2.8. *Let A be a vector space with two bilinear operations $\succ$ and $\prec$. Define a bilinear operation $*$ by Eq. (2.1). Then $(A,\succ,\prec)$ is an anti-associative dendriform algebra if and only if $(A,*)$ is an anti-associative algebra and $(A,L_\succ,R_\prec)$ is a bimodule of $(A,*)$, where two linear maps $L_\succ,R_\prec:A\to\mathrm{End}_{\mathbb{F}}(A)$ are defined by*

$$L_\succ(x)(y)=x\succ y,\quad R_\prec(x)(y)=y\prec x,\quad \forall x,y\in A.$$

Proof. From Eqs. (2.2), (2.3), (2.4) and (2.5) it follows that $(A,\succ,\prec)$ is an anti-associative dendriform algebra if and only if $(A,\succ,\prec)$ is an anti-associative admissible algebra, that is, $(A,*)$ is an anti-associative algebra, and for all $x,y,z\in A$, the following equations hold:

$$x\succ(y\succ z)=-(x\cdot y)\succ z,\ (x\prec y)\prec z=-x\prec(y\cdot z),\ (x\succ y)\prec z=-x\succ(y\prec z).$$

Let $x,y,z\in A$. Then we have

$$\begin{aligned}
L_\succ(x*y)(z)=-L_\succ(x)(L_\succ(y)z) &\iff x\succ(y\succ z)=-(x*y)\succ z,\\
R_\prec(x*y)(z)=-R_\prec(y)(R_\prec(x)z) &\iff (z\prec x)\prec y=-z\prec(x*y),\\
L_\succ(x)(R_\prec(y)z)=-R_\prec(y))(L_\succ(x)z) &\iff x\succ(z\prec y)=-(x\succ z)\prec y.
\end{aligned}$$

Hence $(A,L_\succ,R_\prec)$ is a bimodule. □

Let V be a vector space and V^* be a dual space of V. For $\phi\in V^*$ and $u\in V$, we write $\langle\phi,u\rangle:=\phi(u)$. The following conclusion is obvious.

Lemma 2.9. *Let V be a vector space and V^* be a dual space of V. Let (V,l,r) be a bimodule of an anti-associative algebra A.*

(i) *Let $l^*,r^*:A\to\mathfrak{gl}(V^*)$ be the linear maps given by*

$$\langle l^*(x)u^*,v\rangle=\langle u^*,l(x)v\rangle,\ \langle r^*(x)u^*,v\rangle=\langle u^*,r(x)v\rangle,\quad \forall x\in A,u^*\in V^*,v\in V.$$

Then (V^*,r^*,l^*) is a bimodule of A.

(ii) *$(V,l,0),(V,0,r),(V^*,r^*,0)$ and $(V^*,0,l^*)$ are bimodules of A.*

Proof. (i) Let $x\in A,u^*\in V^*,v\in V$. Then

$$\begin{aligned}
\langle l^*(x*y)u^*,v\rangle=&\langle u^*,l(x*y)v\rangle=-\langle u^*,l(x)(l(y)v)\rangle=-\langle l^*(x)u^*,l(y)v\rangle=-\langle l^*(y)(l^*(x)u^*),v\rangle,\\
\langle r^*(x*y)u^*,v\rangle=&\langle u^*,r(x*y)v\rangle=-\langle u^*,r(y)(r(x)v)\rangle=-\langle r^*(y)u^*,r(x)v\rangle=-\langle r^*(x)(r^*(y)u^*),v\rangle,
\end{aligned}$$

and

$$\begin{aligned}
\langle(l^*(x)r^*(y))u^*,v\rangle=&\langle(l^*(x)(r^*(y)u^*),v\rangle=\langle r^*(y)u^*,l(x)v\rangle=\langle u^*,r(y)(l(x)v)\rangle\\
=&-\langle u^*,l(x)(r(y)v)\rangle=-\langle l^*(x)u^*,r(y)v\rangle=-\langle r^*(y)(l^*(x)u^*),v\rangle,
\end{aligned}$$

Hence, (V^*,r^*,l^*) is a bimodule of A.

(ii) is obvious. □

Corollary 2.10. *Let A be a vector space with two bilinear operations $\succ,\prec:A\otimes A\to A$. Then on the direct sum $\hat{A}:=A\oplus A$ of vector spaces, the following bilinear operation*

$$(x,a)*(y,b)=(x\succ y+x\prec y,x\succ b+a\prec y),\ \ \forall x,y,a,b\in A,\tag{2.6}$$

makes an anti-associative algebra $(\hat{A},*)$ if and only if $(A,\succ,\prec)$ is an anti-associative dendriform algebra.

Proof. The first proof: It is clear that $(\hat{A}, *)$ is an anti-associative algebra if and only if $(A, \succ, \prec)$ is an anti-associative admissible algebra, that is, $(A, *)$ is an associative algebra, and $(A, L_\succ, R_\prec)$ is a bimodule of the associated anti-associative algebra, which is equivalent to the fact that $(A, \succ, \prec)$ is a anti-associative dendriform algebra by Theorem 2.8.

The second proof: By the definition of $\mathcal{C}$-dendriform algebras in the case $\mathcal{C}$ is an anti-associative algebra (see [2]). □

3. anti-associative dendriform algebras and $\mathcal{O}$-operators

M. Aguiar in 2000 [3] was the first who noticed a relation between Rota–Baxter algebras and dendriform algebras. The concept of an $\mathcal{O}$-operator, which generalizes the Rota–Baxter operator, was introduced by C. Bai, L. Guo, and K. Ni in 2010 [4]. They established that any dendriform algebra arises explicitly from an algebra admitting an $\mathcal{O}$-operator. By the motivation of their work, V. Yu. Gubarev and P. S. Kolesnikov proved that every dendriform dialgebra in Var can be embedded into a Rota–Baxter algebra of weight zero in the same variety [22].

Similarly to the above works, the authors of [2] proved the following Proposition.

Proposition 3.1. *Let R be a Rota–Baxter operator on an algebra (A, μ) which belongs to $\mathcal{C}$. For $a, b \in A$, let $a \succ b := R(a)b$ and $a \prec b := aR(b)$. Then $(A, \succ, \prec)$ is a $\mathcal{C}$-dendriform algebra.*

In this section, for the case where $\mathcal{C}$ is an anti-associative algebra, we prove that the result holds more generally for $\mathcal{O}$-operators , which we substitute for the Rota–Baxter operators in the construction.

Definition 3.2. Let $(A, *)$ be an anti-associative algebra and (V, l, r) be a bimodule. A linear map $T : V \to A$ is called an **$\mathcal{O}$-operator** of $(A, *)$ associated to (V, l, r) if the following equation holds:

$$T(u) * T(v) = T\big(l(T(u))v + r(T(v))u\big), \quad \forall u, v \in V.$$

In particular, an $\mathcal{O}$-operator T of $(A, *)$ associated to the bimodule (A, L_*, R_*) is called an **Rota–Baxter operator**, that is, $T : A \to A$ is a linear map satisfying

$$T(x) * T(y) = T\big(T(x) * y + x * T(y)\big), \quad \forall x, y \in A. \tag{3.1}$$

In this case, we call (A, T) a **Rota–Baxter algebra** .

Theorem 3.3. *Let $(A, *)$ be an anti-associative algebra and (V, l, r) be a bimodule. Suppose that $T : V \to A$ is an $\mathcal{O}$-operator of $(A, *)$ associated to (V, l, r). Define two bilinear operations $\succ, \prec$ on V respectively as*

$$u \succ v = l(T(u))v, \quad u \prec v = r(T(v))u, \quad \forall u, v \in V. \tag{3.2}$$

Then $(V, \succ, \prec)$ is an anti-associative dendriform algebra. In this case, T is a homomorphism of anti-associative algebras from the associated anti-associative algebra $(V, *)$ to $(A, *)$. Furthermore, there is an induced anti-associative dendriform algebra structure on $T(V) = \{T(u) \mid u \in V\} \subseteq A$ given by

$$T(u) \succ T(v) = T(u \succ v), \quad T(u) \prec T(v) = T(u \prec v), \quad \forall u, v \in V, \tag{3.3}$$

and T is a homomorphism of anti-associative dendriform algebras.

Proof. Let $u, v, w \in V$. Then we have

$$\begin{aligned} -(u * v) \succ w =& -(u \succ v + u \prec v) \succ w = -l(T\big(l(T(u))v\big))w - l(T\big(r(T(v))u\big))w = -l(T(u) * T(v))w \\ =& l(T(u))(l(T(v))w) = u \succ (v \succ w). \end{aligned}$$

Similarly, we have

$$(u \prec v) \prec w = -u \prec (v * w), \quad (u \succ v) \prec w = -u \succ (v \prec w).$$

Thus by Definition 2.5, $(V, \succ, \prec)$ is an anti-associative dendriform algebra.

Moreover,

$$T(u*v) = T(u\succ v + u\prec v) = T(u\succ v) + T(u\prec v) = T\big(l(T(u))v\big) + T\big(r(T(v))u\big)$$
$$= T(u) * T(v), \quad \forall u, v \in V,$$

shows that T is a homomorphism of anti-associative algebras from the associated anti-associative algebra $(V, *)$ to $(A, *)$

Furthermore, for any $u, v, w \in V$, we can obtain

$$\begin{aligned} T(u) \succ (T(v) \succ T(w)) &= T(u \succ (v \succ w)) = -T(u \succ v) \succ T(w) - T(u \prec v) \succ T(w) \\ &= -(T(u) \succ T(v)) \succ T(w) - (T(u) \prec T(v)) \succ T(w) \\ &= -(T(u) \succ T(v) + T(u) \prec T(v)) \succ T(w). \end{aligned}$$

Similarly, one can derive

$$\begin{aligned} (T(u) \prec T(v)) \prec T(w) &= -T(u) \prec (T(v) \succ T(w) + T(v) \prec T(w)), \\ (T(u) \succ T(v)) \prec T(w) &= -T(u) \succ (T(v) \prec T(w)), \end{aligned}$$

which imply that $T(V)$ is an anti-associative dendriform algebra. This completes the proof. □

Below, we provide an example to illustrate the construction of an anti-associative dendriform algebra using Proposition 3.1.

Example 3.4. Let $(A, *)$ be a complex anti-associative algebra with a basis e_1, e_2 whose non-zero products are given by

$$e_1 * e_1 = e_2.$$

Suppose that $R : A \to A$ is a linear map whose corresponding matrix is given by $\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$ under the basis e_1, e_2. Applying Eq. (3.1) to the pair $\{e_1, e_2\}$ yields $\alpha_{21} = 0$. For the pair $\{e_1, e_1\}$, we obtain $\alpha_{11}(\alpha_{11} - 2\alpha_{22}) = 0$. The other pairs yield no further constraints. Thus R is a Rota–Baxter operator on $(A, *)$ if and only if $\alpha_{11}(\alpha_{11} - 2\alpha_{22}) = 0$. Then we have two cases:

Case 1. $\alpha_{11} = 0$. Then the corresponding matrix of the Rota–Baxter operator is of the form $\begin{pmatrix} 0 & \alpha_{12} \\ 0 & \alpha_{22} \end{pmatrix}$. Hence, one can obtain the trivial anti-associative dendriform algebra.

Case 2. $\alpha_{11} \neq 0$. Then $\alpha_{11} = 2\alpha_{22}$ and the corresponding matrix of the Rota–Baxter operator is of the form $\begin{pmatrix} 2\alpha_{22} & \alpha_{12} \\ 0 & \alpha_{22} \end{pmatrix}$. Hence, we get an anti-associative dendriform algebra whose non-zero products are given by

$$e_1 \succ e_1 = 2\alpha_{22}e_2, \quad e_1 \prec e_1 = 2\alpha_{22}e_2, \quad \alpha_{11} \in \mathbb{C}^*.$$

By the basis change $e_1' = e_1,\ e_2' = 2\alpha_{22}e_2$, it is isomorphic to $Rh_2[1]$ given in Theorem 5.4.

Next we consider invertible $\mathcal{O}$-operators.

Theorem 3.5. *Let $(A, *)$ be an anti-associative algebra. Then there is a compatible anti-associative dendriform algebra structure on $(A, *)$ if and only if there exists an invertible $\mathcal{O}$-operator of $(A, *)$.*

Proof. Suppose that $(A, \succ, \prec)$ is a compatible anti-associative dendriform algebra structure on $(A, *)$. Then

$$x * y = x \succ y + x \prec y = L_\succ(x)y + R_\prec(y)x, \ \ \forall x, y \in A.$$

Hence the identity map $\mathrm{Id} : A \to A$ is an invertible $\mathcal{O}$-operator of $(A, *)$ associated to the bimodule $(A, L_\succ, R_\prec)$.

Conversely, suppose that $T : V \to A$ is an invertible $\mathcal{O}$-operator of $(A, *)$ associated to a bimodule (V, l, r) of $(A, *)$. Then by Proposition 3.3, there exists a compatible anti-associative dendriform

algebra structures on V and $T(V) = A$ defined by Eqs. (3.2) and (3.3) respectively. Let $x, y \in A$. Then there exist $u, v \in V$ such that $x = T(u), y = T(v)$. Hence we have

$$\begin{aligned} x * y &= T(u) * T(v) = T(l(T(u))v + r(T(v))u) = T(u \succ v + u \prec v) \\ &= T(u) \succ T(v) + T(u) \prec T(v) = x \succ y + x \prec y. \end{aligned}$$

So $(A, \succ, \prec)$ is a compatible anti-associative dendriform algebra structure on $(A, *)$. □

Proposition 3.6. *Let $(A, *)$ be an anti-associative algebra and (V, l, r) be a bimodule. Suppose that $T : V \longrightarrow A$ is a linear map. Then T is an $\mathcal{O}$-operator of $(A, *)$ associated to (V, l, r) if and only if the linear map*

$$\hat{T} : A \ltimes_{l,r} V \longrightarrow A \ltimes_{l,r} V, \quad (x, u) \longmapsto (T(u), 0),$$

is a Rota–Baxter operator on the anti-associative algebra $A \ltimes_{l,r} V$.

Proof. Let $x, y \in A, u, v \in V$. Then we have

$$\begin{aligned} \hat{T}((x, u)) * \hat{T}((y, v)) &= (T(u), 0) * (T(v), 0) = (T(u) * T(v), 0), \\ \hat{T}((x, u)) * (y, v) &= (T(u), 0) * (y, v) = (T(u) * y, l(T(u))v), \\ (x, u) * \hat{T}((y, v)) &= (x, u) * (T(v), 0) = (x * T(v), r(T(v))u). \end{aligned}$$

Hence $\hat{T}$ is a Rota–Baxter operator on the anti-associative algebra $A \ltimes_{l,r} V$ if and only if

$$(T(u) * T(v), 0) = \Big(T\big(l(T(u))v + r(T(v))u\big), 0\Big),$$

that is, T is an $\mathcal{O}$-operator of $(A, *)$ associated to (V, l, r). □

Corollary 3.7. *Let $(A, \succ, \prec)$ be a anti-associative dendriform algebra and $(A, *)$ be the associated anti-associative algebra. Set $\hat{A} = A \oplus A$ as the direct sum of vector spaces. Define a bilinear operation $*$ on $\hat{A}$ by Eq. (2.6) and a linear map $\widehat{\mathrm{Id}} : \hat{A} \to \hat{A}$ by*

$$\widehat{\mathrm{Id}}((x, y)) = (y, 0), \ \ \forall x, y \in A.$$

Then $\widehat{\mathrm{Id}}$ is a Rota–Baxter operator on the anti-associative algebra $(\hat{A}, *)$, that is, $(\hat{A}, \widehat{\mathrm{Id}})$ is a Rota–Baxter algebra.

Proof. By Corollary 2.10, $(\hat{A}, *)$ is an associative algebra, which is exactly $A \ltimes_{L_\succ, R_\prec} A$. Since $\mathrm{Id} : A \to A$ is an $\mathcal{O}$-operator of $(A, *)$ associated to the bimodule $(A, L_\succ, R_\prec)$, by Proposition 3.6, $\widehat{\mathrm{Id}}$ is a Rota–Baxter operator on the anti-associative algebra $(\hat{A}, *)$. □

4. anti-associative dendriform algebras and Connes cocycles

In this section, we show that anti-associative dendriform algebras can be obtained from nondegenerate Connes cocycles of anti-associative algebras.

A **Connes cocycle** on an associative algebra $(A, \cdot)$ is an antisymmetric bilinear form $\mathcal{B}$ satisfying

$$\mathcal{B}(x \cdot y, z) + \mathcal{B}(y \cdot z, x) + \mathcal{B}(z \cdot x, y) = 0, \quad \forall x, y, z \in A.$$

It corresponds to the original definition of cyclic cohomology by Connes ([4]). Note that there is a close relation between dendriform algebras and Connes cocycles ([23]). Next we consider a Connes cocycle on anti-associative algebras.

Definition 4.1. Let $(A, *)$ be an anti-associative algebra. A bilinear form $\mathcal{B}$ on A is called a **Connes cocycle** if

$$\mathcal{B}(a * b, c) + \mathcal{B}(b * c, a) + \mathcal{B}(c * a, b) = 0, \quad \forall a, b, c \in A.$$

Theorem 4.2. *Let $(A,*)$ be an anti-associative algebra and $\mathcal{B}$ be a nondegenerate Connes cocycle on $(A,*)$. Then there exists a compatible anti-associative dendriform algebra structure $(A, \succ, \prec)$ on $(A,*)$ defined by*

$$\mathcal{B}(x \succ y, z) = \mathcal{B}(y, z * x), \quad \mathcal{B}(x \prec y, z) = \mathcal{B}(x, y * z), \quad \forall x, y, z \in A. \tag{4.1}$$

Proof. Define a linear map $T : A \to A^*$ by

$$\langle T(x), y\rangle = \mathcal{B}(x, y), \quad \forall x, y \in A,$$

where, A^* is a dual vector space of A. By Lemma 2.9, (A^*, R^*, L^*) is a bimodule of A. Then T is invertible and T^{-1} is an $\mathcal{O}$-operator of the anti-associative algebra $(A,*)$ associated to the bimodule (A^*, R^*, L^*). By Theorem 3.5, there is a compatible anti-associative dendriform algebra structure $\succ, \prec$ on $(A,*)$ given by

$$x \succ y = T^{-1}\big(R^*(x)T(y)\big), \quad x \prec y = T^{-1}\big(L^*(y)T(x)\big), \quad \forall x, y \in A,$$

which gives exactly Eq. (4.1). □

Next, we turn to the double construction of Connes cocycles. Let $(A, *_A)$ be an anti-associative algebra and suppose that there is a anti-associative algebra structure $*_{A^*}$ on its dual space A^*. We construct an anti-associative algebra structure on the direct sum $A \oplus A^*$ of the underlying vector spaces of A and A^* such that both A and A^* are subalgebras and the symmetric bilinear form on $A \oplus A^*$ given by equation

$$\mathcal{B}(x + a^*, y + b^*) = \langle x, b^*\rangle + \langle a^*, y\rangle, \quad \forall x, y \in A, a^*, b^* \in A^*.$$

is a Connes cocycle on $A \oplus A^*$. Such a construction is called a **double construction** of Connes cocycle associated to $(A, *_A)$ and $(A^*, *_{A^*})$ and we denote it by $(T(A) = A \bowtie A^*, \mathcal{B})$.

Proposition 4.3. *Let $(T(A) = A \bowtie A^*, \mathcal{B})$ be a double construction of Connes cocycle. Then there exists a compatible anti-associative dendriform algebra structure $\succ, \prec$ on $T(A)$ defined by equation (4.1). Moreover, A and A^* are anti-associative dendriform subalgebras with this product.*

Proof. The first half follows from Theorem 4.2. Let $x, y \in A$. Set $x \succ y = a + b^*$, where $a \in A$, $b^* \in A^*$. Since A is an anti-associative subalgebra of $T(A)$ and

$$\mathcal{B}(A, A) = \mathcal{B}(A^*, A^*) = 0, \tag{4.2}$$

we have $\mathcal{B}(b^*, A^*) = 0$ and

$$\mathcal{B}(b^*, A) = \mathcal{B}(x \succ y - a, A) \overset{(4.2)}{=} \mathcal{B}(x \succ y, A) \overset{(4.1)}{=} \mathcal{B}(y, A * x) \overset{(4.2)}{=} 0.$$

Therefore $b^* = 0$ due to the nondependence of $\mathcal{B}$. Hence $x \succ y = a \in A$. Similarly, $x \prec y \in A$. Thus A is an anti-associative dendriform subalgebra of $T(A)$ with the product $\succ, \prec$. By symmetry of A and A^*, A^* is also an anti-associative dendriform subalgebra. □

Definition 4.4. Let $(T(A_1) = A_1 \bowtie A_1^*, \mathcal{B}_1)$ and $(T(A_2) = A_2 \bowtie A_2^*, \mathcal{B}_2)$ be two double constructions of Connes cocycles. They are isomorphic if there exists an isomorphism of anti-associative algebras $\phi : T(A_1) \to T(A_2)$ satisfying the following conditions:

$$\phi(A_1) = A_2, \ \phi(A_1^*) = A_2^*, \ \mathcal{B}_1(x, y) = \phi^*\mathcal{B}_2(x, y) = \mathcal{B}_2(\phi(x), \phi(y)), \quad \forall x, y \in A_1. \tag{4.3}$$

Proposition 4.5. *Two double constructions of Connes cocycles $(T(A_1) = A_1 \bowtie A_1^*, \mathcal{B}_1)$ and $(T(A_2) = A_2 \bowtie A_2^*, \mathcal{B}_2)$ are isomorphic if and only if there exists an anti-associative dendriform algebra isomorphism $\phi : T(A_1) \to T(A_2)$ satisfying Eq. (4.3), where the anti-associative dendriform algebra structures on $T(A_1)$ and $T(A_2)$ are given by Eq. (4.1) respectively.*

Proof. It is straightforward. □

5. Isomorphism classes of two dimensional anti-associative dendriform algebras

In this section, we classify two-dimensional anti-associative dendriform algebras. Moreover, we introduce the annihilator of anti-associative dendriform algebras and demonstrate its role in the construction of new algebras from old ones.

Proposition 5.1. *Let $(A, \succ, \prec)$ be an anti-associative dendriform algebra. If one of the binary operations $\succ$, $\prec$ in A vanishes, then with respect to the second one, the algebra A is anti-associative.*

Proof. It is straightforward. □

Definition 5.2. An anti-associative dendriform algebra $(A, \succ, \prec)$ is 2-nilpotent if $(x *_{i_1} y) *_{i_2} z = x *_{i_3} (y *_{i_4} z) = 0$ for all $*_{i_1}, *_{i_2}, *_{i_3}, *_{i_4} \in \{\succ, \prec\}$ and $x, y, z \in A$.

Proposition 5.3. *Let $(A, \succ, \prec)$ be an anti-associative dendriform algebra satisfying $x \succ y + x \prec y = 0$ for all $x, y \in A$. Then $(A, \succ, \prec)$ is a 2-nilpotent algebra.*

Proof. Let $x, y, z \in A$. Then, we obtain

$$\begin{aligned} x \succ (y \succ z) &\overset{(2.3)}{=} -(x \succ y + x \prec y) \succ z = 0, \\ (x \prec y) \prec z &\overset{(2.4)}{=} -x \prec (y \succ z + y \prec z) = 0, \end{aligned}$$

Using these results we can get

$$\begin{aligned} x \prec (y \succ z) &= x \prec (y \succ z) + x \succ (y \succ z) = 0, \\ (x \prec y) \succ z &= (x \prec y) \succ z + (x \prec y) \prec z = 0. \end{aligned}$$

Furthermore, we have

$$\begin{aligned} &(x \succ y) \succ z = (x \succ y) \succ z + (x \prec y) \succ z = (x \succ y + x \prec y) \succ z = 0, \\ &x \prec (y \prec z) = x \prec (y \prec z) + x \prec (y \succ z) = x \prec (y \prec z + y \succ z) = 0, \\ &(x \succ y) \prec z = (x \succ y) \prec z + (x \succ y) \succ z = 0, \\ &x \succ (y \prec z) \overset{(2.5)}{=} (x \succ y) \prec z = 0. \end{aligned}$$

□

Now, we classify anti-associative dendriform algebra structures on 2-dimensional complex vector space.

Theorem 5.4. *Any non-trivial two-dimensional anti-associative dendriform algebra can be included in one of the following classes of algebras:*

Rh_1 : $e_1 \prec e_1 = e_2$

$Rh_2[\lambda]$: $e_1 \succ e_1 = e_2,\ e_1 \prec e_1 = \lambda e_2, \quad \lambda \in \mathbb{C}$.

Proof. Let A be a two dimensional anti-associative dendriform algebra. Set

$$\begin{aligned} e_i \succ e_j &= \alpha_{ij1} e_1 + \alpha_{ij2} e_2, \\ e_i \prec e_j &= \beta_{ij1} e_1 + \beta_{ij2} e_2, \quad 1 \le i, j \le 2. \end{aligned}$$

Now we define $e_i * e_j = e_i \succ e_j + e_i \prec e_j$. Then by Theorem 2.7, the algebra $(A, *)$ is anti-associative. As given in Examples in [8], there are only two non-isomorphic anti-associative algebras i.e. abelian algebra A_1 and the algebra A_2 with non-zero multiplication $e_1 * e_1 = e_2$.

Let the algebra $(A, *)$ be the algebra A_2. Assume that $(A, \succ, \prec)$ is a compatible anti-associative dendriform algebra structure on $(A, \cdot)$. Set $\alpha_{ij} := \alpha_{ij1},\ \beta_{ij} := \alpha_{ij2},\ 1 \le i, j \le 2$.

Then we have

$$\begin{cases} e_1 \prec e_1 = e_2 - \alpha_{11} e_1 - \beta_{11} e_2, \quad e_1 \prec e_2 = -\alpha_{12} e_1 - \beta_{12} e_2, \\ e_2 \prec e_1 = -\alpha_{21} e_1 - \beta_{21} e_2, \quad e_2 \prec e_2 = -\alpha_{22} e_1 - \beta_{22} e_2. \end{cases}$$

Case $\alpha_{22} \neq 0$**.** By considering Eqs. (2.3) and (2.4) for the triple $\{e_2, e_2, e_2\}$ we obtain $\alpha_{21} = -\beta_{22}, \alpha_{12} = -\beta_{22}$, respectively.

Comparing the coefficients of the basis e_1 on the considering Eqs. (2.3) and (2.4) for the triple $\{e_1, e_1, e_1\}$ we obtain $\beta_{22} = 0$.

Again, considering Eq. (2.3) for the triple $\{e_1, e_1, e_2\}$ we get $\alpha_{22} = 0$ which is a contradiction. Hence, there is no such a case.

Case $\alpha_{22} = 0$**.** By considering Eq. (2.3) for the triples $\{e_2, e_2, e_1\}$, $\{e_2, e_2, e_2\}$ and Eq. (2.4) for the triple $\{e_1, e_2, e_2\}$ we obtain $\alpha_{21} = 0, \beta_{22} = 0, \alpha_{12} = 0$, respectively.

Again by considering Eq. (2.3) for the triples $\{e_1, e_1, e_1\}$, $\{e_1, e_1, e_2\}$ and Eq. (2.4) for the triple $\{e_2, e_1, e_1\}$ we obtain $\alpha_{11} = 0, \beta_{12} = 0, \beta_{21} = 0$, respectively. Then we have the anti-associative dendriform algebra with non-zero products given by

$$\begin{cases} e_1 \succ e_1 = \gamma e_2 \\ e_1 \prec e_1 = (1-\gamma)e_2 \end{cases}$$

Then, if $\gamma = 0$ we obtain

$$Rh(A)_1 : \quad e_1 \prec e_1 = e_2,$$

if $\gamma \neq 0$ by the basis change $e_1' = e_1, e_2' = \gamma e_2$ we obtain

$$Rh(A)_2[\lambda] : \quad e_1 \succ e_1 = e_2,\ e_1 \prec e_1 = \lambda e_2, \quad \lambda \in \mathbb{C},\ \lambda \neq -1.$$

Let the algebra $(A, *)$ be abelian algebra A_1. Then by Proposition 5.3 the anti-associative dendriform algebra $(A, \succ, \prec)$ associated to the abelian anti-associative algebra is 2-nilpotent. Thus we have $(x \succ y) \succ z = x \succ (y \succ z) = 0$ which implies that $(A, \succ)$ is an anti-associative 2-nilpotent algebra. Since all 2-dimensional anti-associative algebras are 2-nilpotent, we get $Rh_2[-1]$.

It is possible to consider the multiplication $\prec$ as above. However, it is not difficult to show that the constructed algebras are isomorphic. This finishes the proof. □

Definition 5.5. The following sets are called the annihilator of anti-associative and anti-associative dendriform algebras, respectively:

$$\mathrm{Ann}_{AAs}(A) = \{x \in A \mid x * y = y * x = 0,\ \forall y \in A\},$$

$$\mathrm{Ann}_{AAsD}(A) = \{x \in A \mid x \succ y = x \prec y = y \succ x = y \prec x = 0,\ \forall y \in A\}.$$

An ideal I of an anti-associative dendriform algebra A is a subalgebra of the algebra A that satisfies the conditions:

$$x \succ y,\ x \prec y,\ y \succ x,\ y \prec x \in I, \quad \text{for any} \quad x \in A,\ y \in I.$$

It is obvious that the annihilator of an arbitrary anti-associative dendriform algebra is an ideal.

Now we prove the following Proposition which is a useful tool to construct higher dimensional anti-associative dendriform algebras.

Proposition 5.6. *Let $(A, *)$ be an anti-associative algebra and let $(A, \succ, \prec)$ be a compatible anti-associative dendriform algebra structure on $(A, *)$. If the annihilator $(\mathrm{Ann}_{\mathrm{AAs}}(\mathrm{A}), *)$ of $(A, *)$ is the annihilator $(\mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A}), \succ, \prec)$ of $(A, \succ, \prec)$, then the quotient $(A/\mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A}), \succ, \prec)$ is a compatible anti-associative dendriform algebra structure on $(A/\mathrm{Ann}_{\mathrm{AAs}}(\mathrm{A}), *)$.*

Proof. The Proposition can be described as follows

$$\begin{array}{ccc} (A, \succ, \prec) & \xrightarrow{x*y=x\succ y+x\prec y} & (A, *) \\ {\scriptstyle \bar{x}=x+\mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A})}\downarrow & & \downarrow{\scriptstyle \bar{x}=x+\mathrm{Ann}_{\mathrm{AAs}}(\mathrm{A})} \\ (A/\mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A}), \succ, \prec) & \xrightarrow[\bar{x}*\bar{y}=\bar{x}\succ\bar{y}+\bar{x}\prec\bar{y}]{} & (A/\mathrm{Ann}_{\mathrm{AAs}}(\mathrm{A}), *) \end{array}$$

To show that $(\mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A}), \succ, \prec)$ is a compatible anti-associative dendriform algebra structure on $(A/\mathrm{Ann}_{\mathrm{AAs}}(\mathrm{A}), *)$, it is sufficient to show $\bar{x} * \bar{y} = \overline{x * y}$. The structure is given by $\bar{x} * \bar{y} = \bar{x} \succ \bar{y} + \bar{x} \prec \bar{y}$. Indeed,

$$\begin{aligned} \bar{x} * \bar{y} &= \bar{x} \succ \bar{y} + \bar{x} \prec \bar{y} \\ &= (x + \mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A})) \succ (\mathrm{y} + \mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A})) + (\mathrm{x} + \mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A})) \prec (\mathrm{y} + \mathrm{Ann}_{\mathrm{AAsD}}(\mathrm{A})) \\ &= (x \succ y + x \prec y) + \mathrm{Ann}_{\mathrm{AAs}}(\mathrm{A}) = \overline{\mathrm{x} * \mathrm{y}}. \end{aligned}$$

□

In the following example, we show an application of Proposition 5.6.

Example 5.7. Let $(A, *)$ be an anti-associative algebra with non-zero multiplications: $e_1 * e_2 = -e_2 * e_1 = e_3$. Then $\langle e_3 \rangle$ is the annihilator of the algebra A. Assume $(\langle e_3 \rangle, \prec, \succ)$ is the annihilator of a three-dimensional anti-associative dendriform algebra $(A, \prec, \succ)$ and $(A/\langle e_3 \rangle, \prec, \succ)$ is the algebra $Rh_2[-1]$. Then, according to Proposition 5.6, we can write

$$\begin{cases} e_1 \succ e_1 = e_2 + \alpha_1 e_3, \quad e_1 \succ e_2 = \alpha_2 e_3, \quad e_2 \succ e_1 = \alpha_3 e_3, \quad e_2 \succ e_2 = \alpha_4 e_3 \\ e_1 \prec e_1 = -e_2 - \alpha_1 e_3, \quad e_1 \prec e_2 = (1 - \alpha_2) e_3, \quad e_2 \prec e_1 = -(1 + \alpha_3) e_3, \quad e_2 \prec e_2 = -\alpha_4 e_3. \end{cases}$$

Using Eqs (2.3)-(2.5) for the triples $\{e_1, e_1, e_1\}$ and $\{e_2, e_1, e_1\}$ we obtain $\alpha_2 = \alpha_4 = 0, \alpha_3 = -1$. Now we consider the basis change $e_2' = e_2 + \alpha_1 e_3$. Then we obtain the algebra

$$e_1 \succ e_1 = e_2, \quad e_2 \succ e_1 = -e_3, \quad e_1 \prec e_1 = -e_2, \quad e_1 \prec e_2 = e_3.$$

Acknowledgements

The author would like to express his gratitude to Yunhe Sheng for many fruitful discussions and suggestions in the preparation of this article. The author is very grateful to the referee for useful comments and suggestions.

References

[1] Bai C., Bellier O., Guo L., Ni X., Splitting of Operations, Manin Products, and Rota–Baxter Operators. International Mathematics Research Notices, (2013) Vol. 2013, Iss. 3, P. 485–524.

[2] Ospel C., Panaite F., Vanhaecke P., Polarization and deformations of generalized dendriform algebras. Journal of Noncommutative Geometry, (2022) Vol. 16, P. 561–594.

[3] Aguiar M., Pre-Poisson algebras. Letters in Mathematical Physics, (2000) Vol. 54, P. 263–277.

[4] Chapoton F., Un théorème de Cartier-Milnor-Moore-Quillen pour les bigèbres dendriformes et les algèbres braces. Journal of Pure and Applied Algebra, (2002) Vol. 168, P. 1–18.

[5] Ronco M., Primitive elements in a free dendriform algebra. New Trends in Hopf Algebra Theory. (2000), vol. 267, P. 245–263.

[6] Loday J.-L., Dialgebras. Dialgebras and Related Operads. (2002), vol. 1763, P. 7–66.

[7] Ebrahimi-Fard K., Manchon D., Patras F., New identities in dendriform algebras. Journal of Algebra, (2008) Vol. 320, P. 708–727.

[8] Remm E., Anti-associative algebras. Journal of Algebra and Its Applications, (2024), P. 650080.

[9] Burde D., Fialowski A., Jacobi–Jordan algebras. Linear Algebra and Its Applications, (2014) Vol. 459, P. 586–594.

[10] Kamiya N., Okubo S., Jordan-Lie super algebra and Jordan-Lie triple system. Journal of Algebra, (1997) Vol. 198, P. 388–411.

[11] Zusmanovich P., Special and exceptional mock-Lie algebras. Linear Algebra and Its Applications, (2017) Vol. 518, P. 79–96.

[12] Baklouti A., Benayadi S., Makhlouf A., Mansour S., Jacobi–Jordan-admissible algebras and pre-Jacobi–Jordan algebras. Journal of Algebra and Its Applications, (2025), P. 2550142.

[13] Abdurasulov K., Adashev J., Normatov Z., Solijonova S., Classification of three dimensional anti-dendriform algebras. Communications in Algebra, (2024), P. 1–17.

[14] Bai C., Meng D., The classification of left-symmetric algebras in dimension 2. Chinese Science Bulletin, (1996) Vol. 41, Iss. 23, P. 2207.

[15] Fehlberg Júnior R., Kaygorodov I., Kuster C., The algebraic and geometric classification of antiassociative algebras. RACSAM, (2022) Vol. 116, Iss. 2, P. 78.

[16] Kaygorodov I., Non-associative algebraic structures: classification and structure. Commun. Math. (2024) Vol. 32, Iss. 3, P. 162.

[17] Kobayashi Y., Shirayanagi K., Takahasi S., Tsukada M., A complete classification of three-dimensional algebras over $\mathbb{R}$ and $\mathbb{C}$. Asian-European Journal of Mathematics, (2021) Vol. 14, Iss. 8, P. 2150131.

[18] Petersson H., The classification of two-dimensional nonassociative algebras. Results in Mathematics, (2000) Vol. 37, Iss. 1-2, P. 120–154.

[19] Rikhsiboev I. M., Rakhimov I. S., Basri W., Classification of 3-dimensional complex diassociative algebras. MJMS, (2010) Vol. 4, Iss. 2, P. 241–254.

[20] Rikhsiboev I., Rakhimov I., Basri W., The description of dendriform algebra structures on two-dimensional complex space. JPaNTA, (2010) Vol. 4, Iss. 1, P. 1–18.

[21] Gao D., Liu G., Bai C., Anti-dendriform algebras, new splitting of operations and Novikov type algebras. Journal of Algebraic Combinatorics, (2024) Vol. 59, P. 661–696.

[22] Gubarev V. Y., Kolesnikov P. S., Embedding of dendriform algebras into Rota–Baxter algebras. Central European Journal of Mathematics, (2013) Vol. 11, P. 226–245.

[23] Bai C., Double constructions of Frobenius algebras, Connes cocycles and their duality. Journal of Non-commutative Geometry, (2010) Vol. 4, P. 475–530.

Zafar Normatov,
School of Mathematics, Jilin University, Changchun, China,
V.I. Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences, Tashkent, Uzbekistan
e-mail: z.normatov@inbox.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 187–198
DOI: 10.29229/uzmj.2026-2-22

Asymptotic Properties and Numerical Results of Solutions for a System of Degenerate Parabolic Equations with Nonlinear Boundary Conditions

Rakhmonov Z., Zaripova A.

Abstract. In this article, a system of parabolic equations with nonlinear boundary conditions is transformed into a system of self-similar (automodel) equations by applying a shapeshifting. An asymptotics for a compact supporting solution of the problem is proposed. The obtained asymptotic formulas are used as an initial approximation in the numerical solution of the problem. In numerical experiments $k = 3$ is taken, a finite-difference scheme is proposed, and a Python program is developed to perform computations for various parameter values, corresponding graphs obtained. Numerical experiments have shown that the iterative processes converge within 3 to 5 steps. The compactly supported self-similar asymptotic solution used as the initial approximation played an important role in ensuring the convergence of the iterative process. The graphs illustrate that, for different numerical parameter values, the reaction–diffusion processes occur at a finite speed.

Keywords: System of nonlinear equations, reaction-diffusion, global solution, asymptotic, numerical solution
MSC (2020): 35K55, 35K65, 35K57

1. Introduction

The article considers the following system of nonlinear perturbed equations of parabolic type:

$$u_{it} = \left(|(u_i{}^{m_i})_x|^{p_i-2}(u_i^{m_i})_x\right)_x, \quad i = 1, 2, \ldots, k, \quad x > 0, \quad 0 < t < T, \tag{1.1}$$

with nonlinear boundary conditions

$$-|(u_i{}^{m_i})_x|^{p_i-2}(u_i^{m_i})_x(0,t) = \prod_{j=1}^{k} u_j^{q_{ij}}(0,t), \quad i = 1, 2, \ldots, k\,, \quad 0 < t < T, \tag{1.2}$$

the initial conditions

$$u_i(x, 0) = u_{i0}(x), \quad i = 1, 2, \ldots, k, \quad x > 0. \tag{1.3}$$

We study the self-similar asymptotics and numerical solutions of problem (1.1) - (1.3). Here $k \geq 1$, $p_i > \frac{1}{m_i} - 1$, $m_i \geq 1$, $q_{ij} > 0$ $(i = 1, 2, ..., k)$ are continuous nonnegative functions on $\mathbb{R}_+$.

The system of nonlinear parabolic-type equations (1.1) represents a general mathematical model describing nonlinear heat conduction, diffusion, filtration, chemical reactions, and explosive processes, and is widely used in mathematical modeling of various natural phenomena [1]- [2].

It is known that in the region where $u_i = 0$, the system of equations (1.1) becomes degenerate and has no solution in the classical sense. In such cases, generalized solutions that possess the properties $|\ u_{ix}{}^{m_i}|^{p_i-2}(u_i^{m_i})_x \in C\,(Q)\,,\ 0 \leq u_i \in C\,(Q)$, are considered and this solution satisfies the system of equations (1.1) in the sense of an integral uniqueness [1].

The nonlinear parabolic equations (1.1) originates from the theory of turbulent diffusion [3],[4] (see also the references therein) and appears in population dynamics, chemical reactions, heat transfer, and other related fields. Equation (1.1) encompasses, as particular cases, the porous medium operator (when $p = 0$) and the gradient-diffusive p-Laplacian operator (when $m = 1$), both of which are subjects of intensive study ([5], [6], [7], [8], [9], [10]-[11], [12] and the references therein).

For nonlinear mathematical models, the problems of blow-up and global existence conditions, as well as the study of propagation speeds, have been extensively investigated ([1], [2]- [5], [6], [13], [7], [8]-[14], [15]-[16], [17] and the references therein). In particular, the critical Fujita exponents play an

important role in the initial and boundary value problems for various nonlinear parabolic equations of mathematical physics ([5], [9], [12] and the references therein). The concept of Fujita critical exponents was introduced by Fujita in the 1960s during the study of the heat conduction problem with a nonlinear source term [13].

In the work of Galaktionov and Levine [4], the case of a single equation was studied in the following form:

$$\begin{cases} u_t = (u^m)_{xx,} & x > 0,\ 0 < t < T, \\ -(u^m)_x(0,t) = u^p(0,t), & 0 < t < T, \\ u(x,0) = u_0(x) & x > 0, \end{cases} \tag{1.4}$$

in addition, they also examined the heat conduction equation with gradient diffusion in the following form:

$$\begin{cases} u_t = \left(|u_x|^{m-1}u_x\right)_{x,} & x > 0,\ 0 < t < T, \\ -|u_x|^{m-1}u_x(0,t) = u^p(0,t), & 0 < t < T, \\ u(x,0) = u_0(x) & x > 0. \end{cases} \tag{1.5}$$

Here $m \geq 1,\ p > 0$ and u_0 is a compactly supported nonnegative function (i.e. defined in a bounded domain) serving as the initial function.

They showed that for problem (1.4), the critical exponent for global existence of the solution is $p_0 = \frac{1}{2}(m+1)$ and the critical Fujita exponent is $p_c = m+1$. Moreover, for problem (1.5), the critical exponent for global existence is $p_0 = \frac{2m}{m+1}$ and the critical Fujita exponent is $p_c = 2m$. The critical exponents for global existence and the Fujita critical exponent for problem (1.5) were studied in [7] for the special case $m = 1$.

Xiang, Chen, and Mu [11] investigated the following boundary value problem:

$$\begin{cases} u_{it} = \left(|u_{ix}|^{m_i-1}u_{ix}\right)_x & (i = 1,2,...,k),\ x > 0,\ 0 < t < T, \\ -|u_{ix}|^{m_i-1}u_{ix}(0,t) = u_{i+1}^{p_i}(0,t) & (i = 1,2,...,k),\ u_{k+1} := u_1,\ 0 < t < T, \\ u_i(x,0) = u_{i0}(x) & (i = 1,2,...,k),\ x > 0. \end{cases} \tag{1.6}$$

Here, the parameters satisfy $k \geq 2, m_i > 1, p_i > 0$ and the functions $u_{i0}, (i = 1,2,...,k)$ are continuous and nonnegative. They determined the critical curve of global existence and the Fujita-type critical curve. In article [18], Galaktionov and Levine studied the following equation:

$$u_t = \nabla\left(|\nabla u|^{\sigma}\nabla u^m\right) + u^p, \quad x \in R^N,\ t > 0, u(x,0) = u_0(x),\ x \in R^N,$$

here $\sigma > 0,\ m > 1,\ p > 1$ and $u_0(x)$ is a bounded, positive, and continuous function. They showed that the critical exponent is $p_c = m + \sigma + \frac{\sigma+2}{N}$.

Jiang and Zheng [14] considered the following problem:

$$\begin{cases} u_t = \left(|u_x|^{\beta}(u^m)_x\right)_x, & x > 0,\ 0 < t < T, \\ -|u_x|^{\beta}(u^m)_x(0,t) = u^p(0,t), & 0 < t < T, \\ u(x,0) = u_0(x), & x > 0, \end{cases} \tag{1.7}$$

here $m \geq 1,\ p > 0,\ \beta > 0$. They obtained the critical exponent for global existence $p_0 = \frac{2\beta+m+1}{\beta+2}$ and the critical Fujita exponent $p_c = 2\beta + m + 1$. These results represent a generalized form of the work by Galaktionov and Levine [4].

In [3], the following problem was considered:

$$\begin{cases} u_t = \left(|u_x|^{p_1}\left(u^{m_1}\right)_x\right)_x, & x>0,\ 0<t<T, \\ v_t = \left(|v_x|^{p_2}\left(v^{m_2}\right)_x\right)_x, & x>0, 0<t<T, \\ -|u_x|^{p_1}\left(u^{m_1}\right)_x(0,t) = \alpha_1(0,t)v^{\beta_2}(0,t), & 0<t<T, \\ -|v_x|^{p_2}\left(v^{m_2}\right)_x(0,t) = \alpha_2(0,t)u^{\beta_1}(0,t), & 0<t<T, \\ u(x,0)=u_0(x), \quad v(x,0)=v_0(x), & x>0, \end{cases} \tag{1.8}$$

Here, the parameters satisfy $m_i \geq 1,\ p_i > 0,\ \alpha_i > 0,\ \beta_i > 0\ (i=1,2)$ and the functions $u_0, v_0-$are nonnegative continuous functions with compact support on R_+. They found the global existence critical curve of and the Fujita-type critical curve; however, the classification of the existence and nonexistence of global solutions of the system (1.8) remains a very complex problem.

Using the sources mentioned above, the aim of this article is to find the asymptotic behavior of the solutions to problems (1.1)–(1.3) and based on the asymptotics obtained, to carry out a computational experiment and obtain visualization results for $k=3$.

2. RESEARCH METHODOLOGY

To construct a self-similar system of equations, we seek the solution of the system (1.1) in the following form:

$$\bar{u}_i(x,t) = (\tau+t)^{-k_i} F_i(\xi_i), \quad \xi_i = x(\tau+t)^{-l_i}, \tag{2.1}$$

here $k_i + 1 = (k_i m_i + l_i)(p_i+1) + l_i, \quad (k_i m_i + 1)(p_i+1) = \sum\limits_{j=1}^{k} k_j q_{ij}\ \ (i=1,2,...,k)$.

As a result, of the calculations performed, problem (1.1)–(1.3) takes the following form:

$$\begin{cases} \left(\left|(F_i^{m_i})'\right|^{p_i}(F_i^{m_i})'\right)'(\xi_i) + k_i F_i(\xi_i) + l_i \xi_i F_i'(\xi_i) = 0, \\ -\left|(F_i^{m_i})'\right|^{p_i}(F_i^{m_i})'(0) = \prod\limits_{j=1}^{k} F_j^{q_{ij}}(0), \qquad i=1,2,...,k. \end{cases} \tag{2.2}$$

For problem (2.2), we seek a self-similar solution in the following form:

$$\bar{F}_i(\xi_i) = \left(a_i - b_i \xi_i^{\frac{p_i+2}{p_i+1}}\right)_+^{\frac{p_i+1}{m_i(p_i+1)-1}},$$

here $a_i > 0$, $b_i = \frac{m_i(p_i+1)-1}{(p_i+2)m_i} k_i^{\frac{1}{p_i+1}} > 0,\ \ y_+ = \max(0,y)$.

Theorem 2.1. *For the solution of problem (2.2) with a compact support, as* $\xi_i \to \left(\frac{a_i}{b_i}\right)^{\frac{p_i+1}{p_i+2}}$ *the following asymptotic relation holds*

$$F_i(\xi_i) = \bar{F}_i\left(1 + o\left(1\right)\right). \tag{2.3}$$

Proof: We seek the solution of problem (2.2) in the following form:

$$F_i(\xi_i) = \bar{F}_i(\xi_i) w_i(\tau_i) \tag{2.4}$$

here $\tau_i = -\ln\left(a_i - b_i \xi_i^{\frac{p_i+2}{p_i+1}}\right)$ and $w_i\left(\tau_i\right)-$ are non-negative and bounded functions. Substituting solution (2.4) into problem (2.2), we obtain the following expression:

$$\frac{dF_i^{m_i}}{d\xi_i} = \frac{d\bar{F}_i^{m_i}}{d\tau_i}\frac{d\tau_i}{d\xi_i} w_i{}^{m_i}(\tau_i) + \bar{F}_i^{m_i}(\xi_i)\frac{dw_i{}^{m_i}(\tau_i)}{d\tau_i}\frac{d\tau_i}{d\xi_i} = \frac{d\tau_i}{d\xi_i}\left[\frac{d\bar{F}_i^{m_i}}{d\tau_i} w_i(\tau_i) + \bar{F}_i^{m_i}(\xi_i)\frac{dw_i{}^{m_i}(\tau_i)}{d\tau_i}\right] =$$

$$= \left| \begin{array}{l} \bar{F}_i^{m_i} = e^{-\tau_i \cdot \frac{(p_i+1)m_i}{m_i(p_i+1)-1}} \\ \dfrac{d\bar{F}_i^{m_i}}{d\tau_i} = \dfrac{(p_i+1)m_i}{m_i(p_i+1)-1} e^{-\tau_i \cdot \frac{(p_i+1)m_i}{m_i(p_i+1)-1}} \end{array} \right| = \left| \begin{array}{l} \tau_i = -\ln(a_i - b_i \xi_i^{\frac{p_i+2}{p_i+1}}) \\ \dfrac{d\tau_i}{d\xi_i} = -\dfrac{1}{a_i - b_i \xi_i^{\frac{p_i+2}{p_i+1}}} \left(-b_i \dfrac{p_i+2}{p_i+1} \xi_i^{\frac{1}{p_i+1}} \right) = \\ = \dfrac{b_i \xi_i^{\frac{1}{p_i+1}}}{a_i - b_i \xi_i^{\frac{p_i+2}{p_i+1}}} \dfrac{p_i+2}{p_i+1} = \dfrac{b_i}{e^{-\tau_i}} \dfrac{p_i+2}{p_i+1} \xi_i^{\frac{1}{p_i+1}}; \end{array} \right| =$$

$$= \frac{b_i(p_i+2)e^{\tau_i}}{p_i+1} \xi_i^{\frac{1}{p_i+1}} \left[-\frac{(p_i+1)m_i}{m_i(p_i+1)-1} e^{-\tau_i \frac{(p_i+1)m_i}{m_i(p_i+1)-1}} {w_i}^{m_i}(\tau_i) + e^{-\tau_i \frac{(p_i+1)m_i}{m_i(p_i+1)-1}} ({w_i}^{m_i}(\tau_i))' \right] =$$

$$= -b_i m_i (p_i+2) \xi_i^{\frac{1}{p_i+1}} e^{-\tau_i \frac{1}{m_i(p_i+1)-1}} \left[\frac{({w_i}^{m_i}(\tau_i))'}{(p_i+1)m_i} - \frac{{w_i}^{m_i}(\tau_i)}{m_i(p_i+1)-1} \right],$$

$\frac{d}{d\xi_i} = \frac{d}{d\tau_i} \frac{d\tau_i}{d\xi_i} = -\frac{b_i(p_i+2)e^{\tau_i}}{(p_i+1)} \xi_i^{\frac{1}{p_i+1}} \frac{d}{d\tau_i}$.

Through the found equalities, we obtain the following:

$$\frac{b_i(p_i+2)e^{\tau_i}}{p_i+1} \xi_i^{\frac{1}{p_i+1}} \frac{d}{d\tau_i} \left[(m_i b_i (p_i+2))^{p_i+1} \xi_i e^{-\tau_i \frac{p_i+1}{m_i(p_i+1)-1}} \left[\frac{({w_i}^{m_i}(\tau_i))'}{(p_i+1)m_i} - \frac{{w_i}^{m_i}(\tau_i)}{m_i(p_i+1)-1} \right]^{p_i+1} \right] +$$

$$+ k_i e^{-\tau_i \frac{p_i+1}{m_i(p_i+1)-1}} w_i(\tau_i) + l_i b_i (p_i+2) \xi_i^{\frac{p_i+2}{p_i+1}} e^{-\tau_i \frac{m_i(p_i+1)-2-p_i}{m_i(p_i+1)-1}} \left[\frac{(w_i(\tau_i))'}{(p_i+1)} - \frac{w_i(\tau_i)}{m_i(p_i+1)-1} \right] = 0. \quad (2.5)$$

Let us introduce the following notation:

$$L_1{}^{m_i}(w_i) = \frac{({w_i}^{m_i}(\tau_i))'}{(p_i+1)m_i} - \frac{{w_i}^{m_i}(\tau_i)}{m_i(p_i+1)-1},$$

$$L_2(w_i) = \frac{(w_i(\tau_i))'}{(p_i+1)} - \frac{w_i(\tau_i)}{m_i(p_i+1)-1},$$

as a result

$$-\frac{{b_i}^{p_i+2}(p_i+2){m_i}^{p_i+1}e^{\tau_i}}{p_i+1} \left(a_i - e^{-\tau_i}\right)^{\frac{1}{p_i+2}} b_i^{-1} \frac{d}{d\tau_i} \left[e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} \left(a_i - e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}} [L_1{}^{m_i}(w_i)]^{p_i+1} \right] +$$

$$+ k_i e^{-\tau_i \frac{p_i+1}{m_i(p_i+1)-1}} w_i(\tau_i) + l_i(p_i+2) \left(a_i - e^{-\tau_i}\right) e^{-\tau_i \frac{m_i(p_i+1)-2-p_i}{m_i(p_i+1)-1}} L_2(w_i) = 0. \quad (2.6)$$

From (2.6), we calculate $\frac{d}{d\tau_i}$:

$$\frac{d}{d\tau_i} \left[e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} \left(a_i - e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}} \right] = e^{-\tau_i} \frac{p_i+1}{p_i+2} \left(a_i - e^{-\tau_i}\right)^{-\frac{1}{p_i+2}} e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} -$$

$$- \frac{p_i+1}{m_i(p_i+1)-1} \left(a_i - e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}} e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} =$$

$$= e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} \left(a_i - e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}} \times \left[\frac{p_i+1}{p_i+2} e^{-\tau_i} \left(a_i - e^{-\tau_i}\right)^{-1} - \frac{p_i+1}{m_i(p_i+1)-1} \right],$$

$$\frac{d}{d\tau_i} \left[e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} \left(a_i - e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}} [L_1{}^{m_i}(w_i)]^{p_i+1} \right] = e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i} \left(a_i - e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}} \times$$

$$\times \left[\frac{p_i+1}{p_i+2} e^{-\tau_i} \left(a_i - e^{-\tau_i}\right)^{-1} - \frac{p_i+1}{m_i(p_i+1)-1} \right] [L_1{}^{m_i}(w_i)]^{p_i+1} +$$

$$e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i}\left(a_i-e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}}\frac{d}{d\tau_i}(L_i^{m_i}(\tau_i))^{p_i+1}=e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i}\left(a_i-e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}}\times$$

$$\times\left[\left[\frac{p_i+1}{p_i+2}e^{-\tau_i}\left(a_i-e^{-\tau_i}\right)^{-1}-\frac{p_i+1}{m_i(p_i+1)-1}\right]\left[{L_1}^{m_i}(w_i)\right]^{p_i+1}+\frac{d}{d\tau_i}(L_i^{m_i}(\tau_i))^{p_i+1}\right]=0. \tag{2.7}$$

Substituting (2.7) into (2.6) we obtain :

$$\frac{{b_i}^{p_i+2}(p_i+2)^{p_i+2}{m_i}^{p_i+1}e^{\tau_i}}{p_i+1}\left(a_i-e^{-\tau_i}\right)^{\frac{1}{p_i+2}}b_i^{-1}\left[\left(a_i-e^{-\tau_i}\right)^{\frac{p_i+1}{p_i+2}}e^{-\frac{p_i+1}{m_i(p_i+1)-1}\tau_i}\times\right.$$

$$\left.\times\left[\left[\frac{p_i+1}{p_i+2}e^{-\tau_i}\left(a_i-e^{-\tau_i}\right)^{-1}-\frac{p_i+1}{m_i(p_i+1)-1}\right]\left[{L_1}^{m_i}(w_i)\right]^{p_i+1}+\frac{d}{d\tau_i}(L_i^{m_i}(\tau_i))^{p_i+1}\right]\right]+$$

$$+k_ie^{-\tau_i\frac{p_i+1}{m_i(p_i+1)-1}}w_i(\tau_i)+l_i(p_i+2)\left(a_i-e^{-\tau_i}\right)e^{-\tau_i\frac{m_i(p_i+1)-2-p_i}{m_i(p_i+1)-1}}L_2(w_i)=0,$$

$$\left[\frac{p_i+1}{p_i+2}\frac{e^{-\tau_i}}{\left(a_i-e^{-\tau_i}\right)^{-1}}-\frac{p_i+1}{m_i(p_i+1)-1}\right]\left[{L_1}^{m_i}(w_i)\right]^{p_i+1}+\frac{d}{d\tau_i}(L_i^{m_i}(\tau_i))^{p_i+1}+$$

$$+\frac{k_ie^{-\tau_i}b_i}{t_i\left(a_i-e^{-\tau_i}\right)}w_i(\tau_i)+\frac{l_ib_i(p_i+2)}{t_i}L_2(w_i)=0, \tag{2.8}$$

here $t_i=\frac{{b_i}^{p_i+2}(p_i+2)^{p_i+2}{m_i}^{p_i+1}}{p_i+1}$, $d_1=\frac{p_i+1}{p_i+2}$, $f_1(\tau_i)=\frac{e^{-\tau_i}}{a_i-e^{-\tau_i}}$, $d_2=\frac{k_ib_i}{t_i}$, $d_3=\frac{l_ib_i(p_i+2)}{t_i}$.

Equation (2.8) takes the following form:

$$\left[d_1f_1(\tau_i)-\frac{p_i+1}{m_i(p_i+1)-1}\right]\left[{L_1}^{m_i}(w_i)\right]^{p_i+1}+\frac{d}{d\tau_i}(L_i^{m_i}(\tau_i))^{p_i+1}+d_2f_1(\tau_i)w_i(\tau_i)+d_3L_2(w_i)=0. \tag{2.9}$$

The study of the solutions of the last equation is equivalent to the study of the solutions of equation (2.2), and each of them satisfies the following inequalities on the interval $[\tau_0,+\infty)$

$${w_i}^{m_i}(\tau_i)>0,\quad \frac{{w_i}^{m_i}(\tau_i)}{(p_i+1)m_i}-\frac{\left({w_i}^{m_i}(\tau_i)\right)'}{m_i(p_i+1)-1}\neq 0.$$

Examine whether the solution ${w_i}^{m_i}(\tau_i)$ of equation (2.2) is bounded and has a limit w_0 or becomes unbounded as $\eta\to+\infty$. Suppose that

$$\nu_i(\tau_i)=\left({L_1}^{m_i}(w_i)\right)^{p_i+1}.$$

Then, for the derivative of the function $\nu_i(\tau_i)$, we obtain the following:

$${\nu'}_i=\left[d_1f_1(\tau_i)-\frac{p_i+1}{m_i(p_i+1)-1}\right]\nu_i+d_2f_1(\tau_i)w_i(\tau_i)+d_3L_2(w_i).$$

To analyze the solutions of the last equation, we introduce an auxiliary function:

$$\theta_i(\tau_i,\mu_i)=\left[d_1f_1(\tau_i)-\frac{p_i+1}{m_i(p_i+1)-1}\right]\mu_i+d_2f_1(\tau_i)w_i(\tau_i)+d_3L_2(w_i), \tag{2.10}$$

here $\mu_i-$is a real number. It follows that each function $\theta_i(\tau_i,\mu_i)$is defined in the interval $[\tau_1,+\infty)$ and within this interval, one of the following inequalities holds:

$${\nu'}_i(\tau_i)>0,\quad {\nu'}_i(\tau_i)<0.$$

Therefore, we analyze equation (2.10) taking into account Bol's theorem. It follows that for the function $\nu_i(\tau_i)$ a limit exists for $\tau_i\in[\tau_1,+\infty)$.

Now, let us consider the limiting case. It can be seen in the following:

$$\xi_i\to\left(\frac{a_i}{b_i}\right)^{\frac{p_i+1}{p_i+2}},\ \lim_{\tau_i\to+\infty}f_1(\tau_i)\to 0\quad .$$

Thus, from the existence of the above limit and the facts that $(w_i{}^{m_i})' = 0, \quad (w_i)' = 0$ it follows that

$$\left(d_1 f_1(\tau_i) - \frac{p_i+1}{m_i(p_i+1)-1}\right)\left[\frac{w_i{}^{m_i}(\tau_i)}{(p_i+1)m_i} - \frac{(w_i{}^{m_i}(\tau_i))'}{m_i(p_i+1)-1}\right]^{p_i+1} + \frac{d}{d\tau_i}(L_1{}^{m_i}(w_i))^{p_i+1} +$$

$$+d_2 f_1(\tau_i) w_i(\tau_i) + d_3\left[\frac{(w_i(\tau_i))'}{p_i+1} - \frac{w_i(\tau_i)}{m_i(p_i+1)-1}\right] = 0,$$

$$-\frac{p_i+1}{m_i(p_i+1)-1}\left[\frac{w_i{}^{m_i}(\tau_i)}{(p_i+1)m_i} - \frac{(w_i{}^{m_i}(\tau_i))'}{m_i(p_i+1)-1}\right]^{p_i+1} + d_3\left[\frac{(w_i(\tau_i))'}{p_i+1} - \frac{w_i(\tau_i)}{m_i(p_i+1)-1}\right] = 0.$$

In this case, the following system of algebraic equations with respect to w_i arises:

$$-\frac{p_i+1}{m_i(p_i+1)-1}\left(\frac{w_i(\tau_i)}{m_i(p_i+1)-1}\right)^{p_i+1} = d_2\frac{w_i(\tau_i)}{m_i(p_i+1)-1}.$$

From this system of equations it follows that $w_i = 1$. According to (2.3), $F_i(\xi_i) = \bar{F}_i(1+o(1))$ the theorem is proved.

3. Computational Experiment and Visualization Results

First, we construct a numerical scheme to solve problem (1.1)–(1.3). In the domain $Q = \{(t,x) : 0 \le t < T,\ 0 \le x \le b\}$ we construct a grid in time and space as follows:

$V_\tau = \{t_j = j\tau, \tau > 0, j = 1,2,...,m;\ \ m\tau = T,\ \ T > 0\}$, $S_h = \{x_i = ih; h > 0; i = 1,2,...,n; nh = b\}$ here $\tau = \frac{T}{m}$, and $h = \frac{b}{n}$. To solve problem (1.1)–(1.3) numerically for $k = 3$, we use an implicit scheme.

$$\begin{cases} \dfrac{y_i^{j+1}-y_i^j}{\tau} = \dfrac{1}{h^2}\left[D_{i+1}^j(y_{i+1}{}^j)\left(y_{i+1}^{j+1}-y_i^{j+1}\right) - D_i^j(y_i{}^j)\left(y_i^{j+1}-y_{i-1}^{j+1}\right)\right], \\ \dfrac{z_i^{j+1}-z_i^j}{\tau} = \dfrac{1}{h^2}\left[M_{i+1}^j(z_{i+1}^j)\left(z_{i+1}^{j+1}-z_i^{j+1}\right) - M_i^j(z_i^j)\left(z_i^{j+1}-z_{i-1}^{j+1}\right)\right], \\ \dfrac{t_i^{j+1}-t_i^j}{\tau} = \dfrac{1}{h^2}\left[N_{i+1}^j(t_{i+1}^j)\left(t_{i+1}^{j+1}-t_i^{j+1}\right) - N_i^j(t_i^j)\left(t_i^{j+1}-t_{i-1}^{j+1}\right)\right], \end{cases} \tag{3.1}$$

$$i = 2,...,n-1, j = 0,1,...,m-1,$$

$$y_i^0 = y(x_i,0),\ z_i^0 = z(x_i,0),\ t_i^0 = t(x_i,0),$$
$$i = 0,1,...,n,$$

$$\begin{cases} -D_1(y_1)\cdot\dfrac{\left(y_1^{j+1}\right)^{m_0} - \left(y_0^{j+1}\right)^{m_0}}{h} = (u_j)^{q_{1j}}, \\ -M_1(z_1)\cdot\dfrac{\left(z_1^{j+1}\right)^{m_0} - \left(z_0^{j+1}\right)^{m_0}}{h} = (u_j)^{q_{2j}}, \\ -N_1(t_1)\cdot\dfrac{\left(t_1^{j+1}\right)^{m_0} - \left(t_0^{j+1}\right)^{m_0}}{h} = (u_j)^{q_{3j}}, \end{cases} \tag{3.2}$$

$$\begin{cases} y_n^j = \varphi_1(k_j), \\ z_n^j = \varphi_2(k_j), \quad j = 2,3,...,m \\ t_n^j = \varphi_3(k_j) \end{cases} \tag{3.3}$$

a)

$$\begin{cases} D_i^j(y_i^j) = D_i^j\left(\dfrac{y_i^j - y_{i-1}^j}{2}\right), \\ M_i^j(z_i^j) = M_i^j\left(\dfrac{z_i^j - z_{i-1}^j}{2}\right), \\ N_i^j(t_i^j) = N_i^j\left(\dfrac{t_i^j - t_{i-1}^j}{2}\right), \end{cases} \tag{3.4}$$

b)

$$\begin{cases} D_i^j\left(y_i^j\right) = \dfrac{D_i^j(y_i^j) - D_{i-1}^j(y_{i-1}^j)}{2}, \\ M_i^j\left(z_i^j\right) = \dfrac{M_i^j(z_i^j) - M_{i-1}^j(z_{i-1}^j)}{2}, \\ N_i^j\left(t_i^j\right) = \dfrac{N_i^j(t_i^j) - N_{i-1}^j(t_{i-1}^j)}{2}, \end{cases} \tag{3.5}$$

here, equations (3.4)–(3.5) are written using the Samarskii–Sobol scheme. If we apply formula (3.4), we obtain

$$\begin{cases} D_i^j(y_i^j) = m_i \dfrac{\left(y_i^j\right)^{m_i-1} + \left(y_{i-1}^j\right)^{m_i-1}}{2} \left| \dfrac{\left(y_i^j\right)^{m_i} - \left(y_{i-1}^j\right)^{m_i}}{h} \right|^{p_1}, \\ D_{i+1}^j(y_{i+1}^j) = m_i \dfrac{\left(y_i^j\right)^{m_i-1} + \left(y_{i+1}^j\right)^{m_i-1}}{2} \left| \dfrac{\left(y_i^j\right)^{m_i} - \left(y_{i-1}^j\right)^{m_i}}{h} \right|^{p_1}, \end{cases}$$

$$\begin{cases} M_i^j(z_i^j) = m_i \dfrac{\left(z_i^j\right)^{m_i-1} + \left(z_{i-1}^j\right)^{m_i-1}}{2} \left| \dfrac{\left(z_i^j\right)^{m_i} - \left(z_{i-1}^j\right)^{m_i}}{h} \right|^{p_2}, \\ M_{i+1}^j(z_{i+1}^j) = m_i \dfrac{\left(z_i^j\right)^{m_i-1} + \left(z_{i+1}^j\right)^{m_i-1}}{2} \left| \dfrac{\left(z_{i+1}^j\right)^{m_i} - \left(z_i^j\right)^{m_i}}{h} \right|^{p_2}, \end{cases} \tag{3.6}$$

$$\begin{cases} N_i^j(t_i^j) = m_i \dfrac{\left(t_i^j\right)^{m_i-1} + \left(t_{i-1}^j\right)^{m_i-1}}{2} \left| \dfrac{\left(t_i^j\right)^{m_i} - \left(t_{i-1}^j\right)^{m_i}}{h} \right|^{p_3}, \\ N_{i+1}^j(t_{i+1}^j) = m_i \dfrac{\left(t_i^j\right)^{m_i-1} + \left(t_{i+1}^j\right)^{m_i-1}}{2} \left| \dfrac{\left(t_{i+1}^j\right)^{m_i} - \left(t_i^j\right)^{m_i}}{h} \right|^{p_3}, \end{cases}$$

if we apply formula (3.6), we obtain:

$$\begin{cases} D_i^j(y_i^j) = m_i \dfrac{\left(y_i^j\right)^{m_i-1} + \left(y_{i-1}^j\right)^{m_i-1}}{2} \dfrac{\left|\frac{y_i^j - y_{i-1}^j}{h}\right|^{p_1} + \left|\frac{y_{i-1}^j - y_{i-2}^j}{h}\right|^{p_1}}{2}, \\ D_{i+1}^j(y_{i+1}^j) = m_i \dfrac{\left(y_i^j\right)^{m_i-1} + \left(y_{i+1}^j\right)^{m_i-1}}{2} \dfrac{\left|\frac{y_{i+1}^j - y_i^j}{h}\right|^{p_1} + \left|\frac{y_i^j - y_{i-1}^j}{h}\right|^{p_1}}{2}, \end{cases}$$

$$\begin{cases} M_i^j(z_i^j) = m_i \dfrac{\left(z_i^j\right)^{m_i-1} + \left(z_{i-1}^j\right)^{m_i-1}}{2} \dfrac{\left|\frac{z_i^j - z_{i-1}^j}{h}\right|^{p_2} + \left|\frac{z_{i-1}^j - z_{i-2}^j}{h}\right|^{p_2}}{2}, \\ M_{i+1}^j(z_{i+1}^j) = m_i \dfrac{\left(z_i^j\right)^{m_i-1} + \left(z_{i+1}^j\right)^{m_i-1}}{2} \dfrac{\left|\frac{z_{i+1}^j - z_i^j}{h}\right|^{p_2} + \left|\frac{z_i^j - z_{i-1}^j}{h}\right|^{p_2}}{2}, \end{cases} \tag{3.7}$$

$$\begin{cases} N_i^j(t_i^j) = m_i \dfrac{\left(t_i^j\right)^{m_i-1} + \left(t_{i-1}^j\right)^{m_i-1}}{2} \dfrac{\left|\frac{t_i^j - t_{i-1}^j}{h}\right|^{p_3} + \left|\frac{t_{i-1}^j - t_{i-2}^j}{h}\right|^{p_3}}{2}, \\ N_{i+1}^j(z_{i+1}^j) = m_i \dfrac{\left(t_i^j\right)^{m_i-1} + \left(t_{i+1}^j\right)^{m_i-1}}{2} \dfrac{\left|\frac{t_{i+1}^j - t_i^j}{h}\right|^{p_3} + \left|\frac{t_i^j - t_{i-1}^j}{h}\right|^{p_3}}{2}, \end{cases}$$

system (3.1) is a system of nonlinear equations with respect to y^{j+1}, z^{j+1}, and t^{j+1}. To obtain numerical solutions of this system, we use the simple iteration method to linearize it.

$$\begin{cases} \dfrac{\overset{s+1}{y}{}_i^{j+1} - \overset{s}{y}{}_i^j}{\tau} = \dfrac{1}{h^2}\left[D_{i+1}^j(y_{i+1}^j)\left(\overset{s+1}{y}{}_{i+1}^{j+1} - \overset{s+1}{y}{}_i^{j+1}\right) - D_i^j(y_i^j)\left(\overset{s+1}{y}{}_i^{j+1} - \overset{s+1}{y}{}_{i-1}^{j+1}\right)\right], \\ \dfrac{\overset{s+1}{z}{}_i^{j+1} - \overset{s}{z}{}_i^j}{\tau} = \dfrac{1}{h^2}\left[M_{i+1}^j(z_{i+1}^j)\left(\overset{s+1}{z}{}_{i+1}^{j+1} - \overset{s+1}{z}{}_i^{j+1}\right) - M_i^j(z_i^j)\left(\overset{s+1}{z}{}_i^{j+1} - \overset{s+1}{z}{}_{i-1}^{j+1}\right)\right], \\ \dfrac{\overset{s+1}{t}{}_i^{j+1} - \overset{s}{t}{}_i^j}{\tau} = \dfrac{1}{h^2}\left[N_{i+1}^j(t_{i+1}^j)\left(\overset{s+1}{t}{}_{i+1}^{j+1} - \overset{s+1}{t}{}_i^{j+1}\right) - N_i^j(t_i^j)\left(\overset{s+1}{t}{}_i^{j+1} - \overset{s+1}{z}{}_{i-1}^{j+1}\right)\right], \end{cases} \tag{3.8}$$

here, $s = 0, 1, 2, ...$

The performance of the iterative method is highly dependent on the selection of a suitable initial approximation. A well-chosen approximation ensures rapid convergence to the solution while preserving the physical meaning of the problem. A fitting choice for this initial approximation is the asymptotic solution of the problem.

For each time step, the initial iterative values $\overset{s+1}{y}{}_i$, $\overset{s+1}{z}{}_i$, $\overset{s+1}{t}{}_i$ are taken from the previous time step: $\overset{0}{y}{}^{j+1} = y^j$, $\overset{0}{z}{}^{j+1} = z^j, \overset{0}{t}{}^{j+1} = t^j$. When using an iterative scheme for calculations, the precision of the iteration (repetition process) is determined.

$$\begin{cases} \max\limits_{0\le i\le n}\left|\overset{s+1}{y}{}_i - \overset{s}{y}{}_i\right| < \varepsilon, \\ \max\limits_{0\le i\le n}\left|\overset{s+1}{z}{}_i - \overset{s}{z}{}_i\right| < \varepsilon, \\ \max\limits_{0\le i\le n}\left|\overset{s+1}{t}{}_i - \overset{s}{t}{}_i\right| < \varepsilon. \end{cases} \tag{3.9}$$

Let us introduce the following notation: $\overset{-}{y}{}_i = y_i^{j+1}$, $\overset{-}{z}{}_i = z_i^{j+1}$, $\overset{-}{t}{}_i = t_i^{j+1}$. Then the systems (3.8) can be written in the following form.

$$\begin{cases} \overset{s}{A}_{1i}\overset{s+1}{\overline{y}}{}_{i-1} - \overset{s}{C}_{1i}\overset{s+1}{\overline{y}}{}_i + \overset{s}{B}_{1i}\overset{s+1}{\overline{y}}{}_{i+1} = -\overset{s}{F}_{1i}, \\ \overset{s}{A}_{2i}\overset{s+1}{\overline{z}}{}_{i-1} - \overset{s}{C}_{2i}\overset{s+1}{\overline{z}}{}_i + \overset{s}{B}_{2i}\overset{s+1}{\overline{z}}{}_{i+1} = -\overset{s}{F}_{2i}, \\ \overset{s}{A}_{3i}\overset{s+1}{\overline{z}}{}_{i-1} - \overset{s}{C}_{3i}\overset{s+1}{\overline{z}}{}_i + \overset{s}{B}_{3i}\overset{s+1}{\overline{z}}{}_{i+1} = -\overset{s}{F}_{3i}, \end{cases} \tag{3.10}$$

here,

$$\overset{s}{A}_{1i} = \frac{\tau}{h^2}D_i^j(\overset{s}{y}{}_i^j), \overset{s}{A}_{2i} = \frac{\tau}{h^2}M_i^j(\overset{s}{z}{}_i^j), \overset{s}{A}_{3i} = \frac{\tau}{h^2}N_i^j(\overset{s}{t}{}_i^j),$$

$$\overset{s}{B}_{1i} = \frac{\tau}{h^2} D^j_{i+1}(\overset{s}{y^j_{i+1}}), \overset{s}{B}_{2i} = \frac{\tau}{h^2} M^j_{i+1}(\overset{s}{z^j_{i+1}}), \overset{s}{B}_{3i} = \frac{\tau}{h^2} N^j_{i+1}(\overset{s}{t^j_{i+1}}),$$

$$\overset{s}{C}_{1i} = \frac{\tau}{h^2}\left(D^j_i\left(\overset{s}{y^j_i}\right) + D^j_{i+1}\left(\overset{s}{y^j_{i+1}}\right)\right) + 1, \overset{s}{C}_{2i} = \frac{\tau}{h^2}\left(M^j_i\left(\overset{s}{z^j_i}\right) + M^j_{i+1}\left(\overset{s}{z^j_{i+1}}\right)\right) + 1,$$

$$\overset{s}{C}_{3i} = \frac{\tau}{h^2}\left(N^j_i\left(\overset{s}{t^j_i}\right) + N^j_{i+1}\left(\overset{s}{t^j_{i+1}}\right)\right) + 1,$$

$$\overset{s}{F}_{1i} = \overset{s}{y^j_i}, \overset{s}{F}_{2i} = \overset{s}{z^j_i}, \overset{s}{F}_{3i} = \overset{s}{t^j_i}.$$

The required values at the end of the interval $0 \le x \le b$ can be obtained using Milne's second order precision formula.

$$\begin{aligned}
&\left.\frac{\partial y}{\partial x}\right|_0 \approx \frac{-y_2 + 4y_1 - 3y_0}{2h}, \quad \left.\frac{\partial y}{\partial x}\right|_n \approx \frac{3y_n - 4y_{n-1} + y_{n-2}}{2h}, \\
&\left.\frac{\partial z}{\partial x}\right|_0 \approx \frac{-z_2 + 4z_1 - 3z_0}{2h}, \quad \left.\frac{\partial z}{\partial x}\right|_n \approx \frac{3z_n - 4z_{n-1} + z_{n-2}}{2h}, \\
&\left.\frac{\partial t}{\partial x}\right|_0 \approx \frac{-t_2 + 4t_1 - 3t_0}{2h}, \quad \left.\frac{\partial t}{\partial x}\right|_n \approx \frac{3t_n - 4t_{n-1} + t_{n-2}}{2h}\ .
\end{aligned} \tag{3.11}$$

To obtain the numerical solution of the linear algebraic equation system (3.10), we use the sweep method. According to the sweep method:

$$\begin{cases}
\bar{y}_i = \alpha_{1i}\bar{y}_{i+1} + \beta_{1i}, \\
\bar{z}_i = \alpha_{2i}\bar{z}_{i+1} + \beta_{2i}, \\
\bar{t}_i = \alpha_{3i}\bar{t}_{i+1} + \beta_{3i},
\end{cases} \tag{3.12}$$

here, $\alpha_{1i}, \alpha_{2i}, \alpha_{3i}, \beta_{1i}, \beta_{2i}, \beta_{3i}$ are the coefficients that are calculated using the following formulas:

$$\begin{cases}
\alpha_{1,i+1} = \dfrac{B_{1i}}{C_{1i} - \alpha_{1i}A_{1i}} \\
\beta_{1,i+1} = \dfrac{A_{1i}\beta_{1i} + F_{1i}}{C_{1i} - \alpha_{1i}A_{1i}}
\end{cases}$$

$$\begin{cases}
\alpha_{2,i+1} = \dfrac{B_{2i}}{C_{2i} - \alpha_{2i}A_{2i}} \\
\beta_{2,i+1} = \dfrac{A_{2i}\beta_{2i} + F_{2i}}{C_{2i} - \alpha_{2i}A_{2i}}
\end{cases}$$

$$\begin{cases}
\alpha_{3,i+1} = \dfrac{B_{3i}}{C_{3i} - \alpha_{3i}A_{3i}} \\
\beta_{3,i+1} = \dfrac{A_{3i}\beta_{3i} + F_{3i}}{C_{3i} - \alpha_{3i}A_{3i}}
\end{cases}$$

here, $i = 1, 2, ..., n$ and the values $\alpha_{10}, \alpha_{20}, \alpha_{30}, \beta_{10}, \beta_{20}, \beta_{30}$ are determined from the boundary conditions (3.12).

A computational experiment was carried out using the numerical scheme constructed above. We present some of its results. For calculations, the grid step was taken as $h = 0.05$ the number of nodes $N = 300$ and the iteration accuracy was set to $\varepsilon = 10^{-3}$. Formulas (2.1) and (2.3) were used as an initial approximation for the iterative process.

□

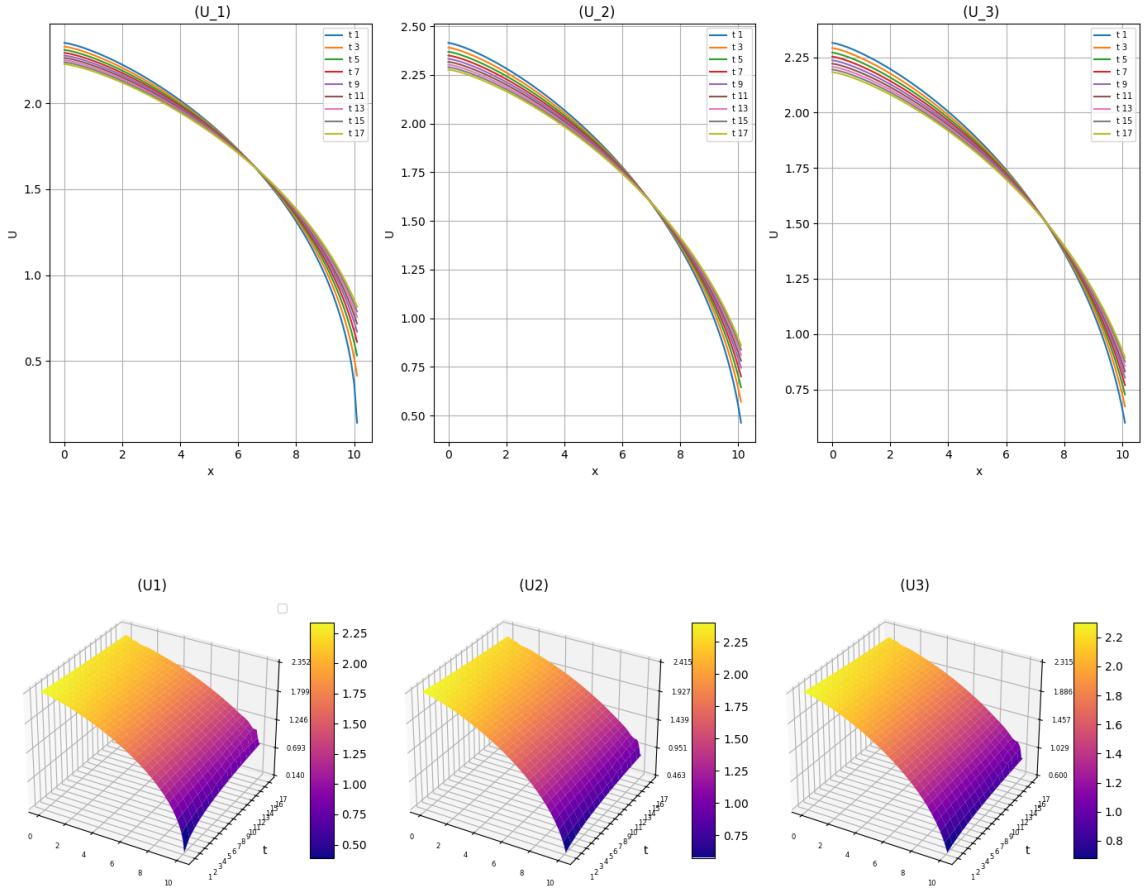


FIGURE 1. Numerical solution of problem (1.1) - (1.3).

q_{11}=4.61, q_{12}=4.21, q_{13}=4.42, q_{21}=4.32, q_{22}=4.51, q_{23}=4.61
q_{31}=4.33, q_{32}=4.63, q_{33}=4.83, p_1=2.18, p_2=2.2, p_3=2.21
m_1=2.59, m_2=2.52, m_3=2.63

Figure 1,2 shows the results of the numerical solution of the problem (1.1) - (1.3) for $p_i > \frac{1}{m_i} - 1$, corresponding to the case of slow diffusion. (2.1),(2.3) were taken as the initial approximation for the iterative process. At $p_i > \frac{1}{m_i} - 1$, as follows from the asymptotic formulasv(2.3) of the graphs, the heat propagation occurs at a finite rate. The penetration depth of a thermal wave depends on the time and the wavefront (the point at which $u_i(x,t)$ becomes zero) for each medium located at the final point: $x_{\phi_{u_i}} = \left(\frac{a_i}{b_i}\right)^{\frac{p_i+1}{p_i+2}} (\tau + t)^{l_i} < \infty$.

Numerical experiments have shown that iterative processes converge in 3 to 5 steps. The asymptotics of the compactly supported self-similar solutions, taken as the initial approximation, played an important role in ensuring the convergence of the iterative process. In the graphs presented above, it can be seen that for the given numerical parameter values, the reaction-diffusion processes occur at a finite speed.

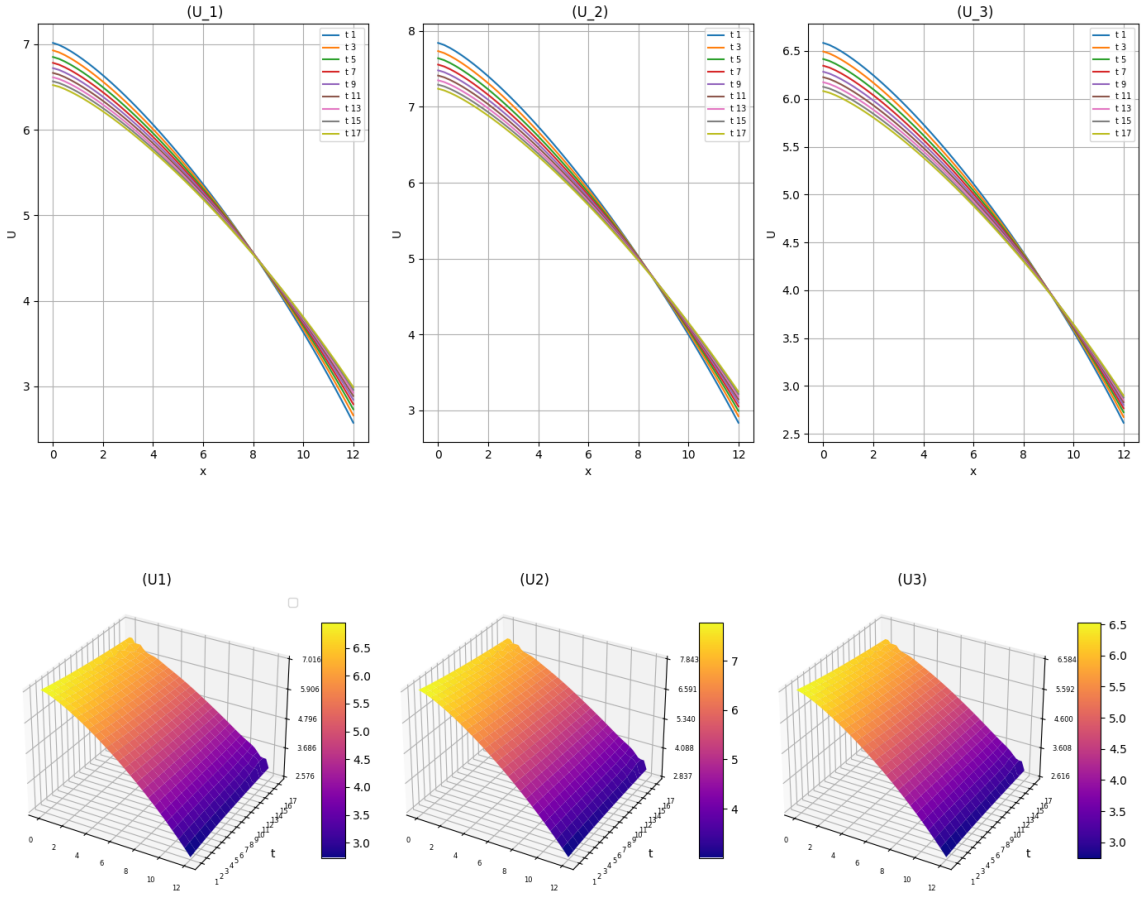


FIGURE 2. Numerical solution of problem (1.1) - (1.3).

q_{11}=3.601, q_{12}=3.201, q_{13}=3.402, q_{21}=3.302, q_{22}=3.501, q_{23}=3.601
q_{31}=3.303, q_{32}=3.603, q_{33}=3.803, p_1=2.18, p_2=2.19, p_3=2.21
m_1=1.59, m_2=1.52, m_3=1.63.

4. Conclusion

It is well known that in the numerical solution of nonlinear problems, one of the main issues is choosing an initial approximation that ensures rapid convergence of the iterative process to the exact solution and preserving the qualitative properties of nonlinear processes. This problem is solved by using the asymptotic formulas constructed above as the initial approximation, corresponding to the values of the numerical parameters. The results of several computational experiments for $k = 3$, along with the obtained graphs and the analysis of numerical solutions, are presented. The outcomes of the computational experiments demonstrated the efficiency of the proposed methods, while the numerical solutions reflected the nonlinear characteristics of the process under consideration.

References

[1] Astretta G., Marrucci G., Principles of Non-Newtonian Fluid Mechanics. McGraw-Hill, New York, (1974).

[2] Chen B. T., Mi Y. S., Mu C. L., Critical exponents for a doubly degenerate parabolic system coupled via nonlinear boundary flux. Acta Mathematica Scientia, (2011) Vol. 31B, P. 681–693.

[3] Cui Z. J., Critical curves of the non-Newtonian polytropic filtration equations coupled with nonlinear boundary conditions. Nonlinear Analysis, (2008) Vol. 68, P. 3201–3208.

[4] Ivanov A. V., Hoder estimates for quasilinear doubly degenerate parabolic equations. Journal of Soviet Mathematics, (1991) Vol. 56, P. 2320–2347.

[5] Deng K., Levine H. A., The role of critical exponents in blow-up theorems: The sequel. Journal of Mathematical Analysis and Applications, (2000) Vol. 243, P. 85–126.

[6] Ferreira R., Pablo A., Quiros F., Rossi J. D., The blow-up profile for a fast diffusion equation with a nonlinear boundary condition. Rocky Mountain Journal of Mathematics, (2003) Vol. 33, P. 123–146.

[7] Galaktionov V. A., Levine H. A., On critical Fujita exponents for heat equations with nonlinear flux conditions on the boundary. Israel Journal of Mathematics, (1996) Vol. 94, P. 125–146.

[8] Jiang Z. X. Doubly degenerate parabolic equation with nonlinear inner sources or boundary flux. PhD thesis. China: Dalian University of Technology, 2009.

[9] Levine H. A., The role of critical exponents in blow up theorems. SIAM Review, (1990) Vol. 32, P. 262–288.

[10] Mi Y. S., Mu C. L., Chen B. T. Blow-up analysis for a fast diffusive parabolic equation with nonlinear boundary flux. preprint.

[11] Mi Y. S., Mu C. L., Chen B. T. Critical exponents for a fast diffusive parabolic system coupled via nonlinear boundary flux. preprint.

[12] Vazquez J. L., The Porous Medium Equations: Mathematical Theory. Oxford University Press, Oxford, UK, (2007).

[13] Fujita H., On the blowing up of solutions of the Cauchy problem $u_t = \Delta u + u^{1+\alpha}$. Journal of the Faculty of Science, University of Tokyo, Section I, (1966) Vol. 13, P. 109–124.

[14] Kalashnikov A. S., Some problems of the qualitative theory of nonlinear degenerate parabolic equations of second order. Uspekhi Mat. Nauk. (1987) Vol. 42, P. 135–176.

[15] Li Z. P., Mu C. L., Critical curves for fast diffusive polytropic filtration equation coupled via nonlinear boundary flux. Journal of Mathematical Analysis and Applications, (2008) Vol. 346, P. 55–64.

[16] Li Z. P., Mu C. L., Cui Z. J., Critical curves for a fast diffusive polytropic filtration system coupled via nonlinear boundary flux. Zeitschrift für Angewandte Mathematik und Physik, (2009) Vol. 60, P. 284–296.

[17] Quiros F., Rossi J. D., Blow-up set and Fujita-type curves for a degenerate parabolic system with nonlinear conditions. Indiana University Mathematics Journal, (2001) Vol. 50, P. 629–654.

[18] Galaktionov V. A., Levine H. A., A general approach to critical Fujita exponents and systems. Nonlinear Analysis, (1998) Vol. 34, P. 1005–1027.

Rakhmonov Zafar Ravshanovich,
National University of Uzbekistan, Tashkent, Uzbekistan
e-mail: zraxmonov@inbox.ru

Zaripova Aziza Rakhmonovna,
Karshi State University, Karshi, Uzbekistan
e-mail: azizazaripova1990@mail.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 199–207
DOI: 10.29229/uzmj.2026-2-23

Application of the method of quadrature's with derivatives for solving problems of the Euler - Bernoulli beam

Usmanov Kh.I.

Abstract. This paper proposes a method for solving the problem of the Euler - Bernoulli beam, in which the problem is reduced to the solution of the Fredholm integral equation of the second kind. This integral equation is solved by the method of optimal quadrature formulas with derivatives. Then, using the Green function, the analytic solution is determined. Different boundary conditions can be used to determine the numerical and analytical solutions. To determine the suitability of the method, the numerical results are compared with the analytical solution.

Keywords: differential equation, integral equation, optimal quadrature formula, beam deflection value, absolute error
MSC (2020): 53C12, 57R25, 57R35

1. Introduction

In connection with the development of high-speed rail transportation, the problem of modeling the dynamic effects of a rolling load on bending elements is widely used in the field of design and control of the safety of transport facilities. This issue is increasingly relevant in relation to the increase in traffic speeds, especially in the field of railway transport. The theoretical foundations of the control complex method are laid out in the works [1] - [2].

Mathematical models and the calculation of structures that lie on an elastic base are the subject of many research papers [3] - [4]. In these works, mathematical models of Timoshenko, Rayleigh, and Euler-Bernoulli were used. To solve the problems, the Fourier integral transform method, the finite element method, and the finite difference method were used.

The development of numerical methods, in particular, the widespread use of the optimal quadrature formula method, made it possible to model various structural systems in sufficient detail under the influence of various external influences.

This paper proposes a new approach to solving the problem of the reliability of railway tracks. The essence of this approach lies in the fact that the Euler-Bernoulli model is chosen for the main calculation model, the differential equation is reduced to the Fredholm integral equation of the second kind, which is solved by the method of optimal quadrature formulas with derivatives.

An analytical solution of this model is given. A specific example is considered. The numerical results are compared with the analytical solution. Appropriate conclusions are issued on the method.

2. Problem Statement

We consider the following problem: a beam of length L, lying on a solid elastic base, under the action of a moving load with P a constant velocity V. You must define the deflection of the beam at each point and in each time layer.

To solve this problem, the following mathematical models of beam oscillation can be considered [3]:
Tymoshenko's model

$$EI\frac{\partial^4 u}{\partial x^4} - \rho I\frac{\partial^4 u}{\partial x^2\partial t^2} - \frac{\rho EI}{\gamma G}\frac{\partial^4 u}{\partial x^2\partial t^2} + \frac{\rho I}{\gamma G}\frac{\partial^4 u}{\partial t^4} + +\rho A\frac{\partial^2 u}{\partial t^2} + c\frac{\partial u}{\partial t} + ku = q(x,t), \tag{1}$$

Rayleigh model

$$EI\frac{\partial^4 u}{\partial x^4} - \rho I\frac{\partial^4 u}{\partial x^2\partial t^2} + \rho A\frac{\partial^2 u}{\partial t^2} + c\frac{\partial u}{\partial t} + ku = q(x,t), \tag{2}$$

Euler-Bernoulli model

$$EI\frac{\partial^4 u}{\partial x^4} + \rho A\frac{\partial^2 u}{\partial t^2} + c\frac{\partial u}{\partial t} + ku = q(x,t), \tag{3}$$

where E is the modulus of elasticity of the beam; G is modulus of shear; I is the moment of inertia of the beam cross-section; ρ is beam material density; A is cross-sectional area of the beam; k is coefficient of the bed of the elastic base; c is base damping coefficient; γ is the coefficient of influence of the form; $q(x,t)$ is the intensity of the impact on the beam.

Conductive rails mainly hold shear forces and the transverse dimensions of the rail are small compared to its length, so that the rails can be considered within the framework of the Euler-Bernoulli beam model. The intensity of the impact on the beam is of a harmonic type. Therefore, imagine a moving load

$$q(x,t) = P\cos\omega t\delta(x-Vt),$$

where ω is the vibration frequency of the beam.

As the main calculation model, let us consider the differential equation of the Euler-Bernoulli beam (the basis of the Winkler model):

$$EI\frac{\partial^4 u}{\partial x^4} + \rho A\frac{\partial^2 u}{\partial t^2} + ku = P\cos\omega t\delta(x-Vt), x\in[0,L], t\in[0,T], \tag{3}$$

with the corresponding initial conditions:

$$u(x,0) = 0, \frac{\partial u(x,0)}{\partial t} = 0, x\in[0,L],$$

where L is the length of the region under consideration, T is the study period,

$$\delta(x-Vt) = \begin{cases} 1, & \text{if } x = Vt, \\ 0, & \text{if } x \neq Vt. \end{cases}$$

As boundary conditions, we consider different models of problems:

Model 1. Beam with fixed ends:

$$u(0,t) = 0, \frac{\partial u(0,t)}{\partial x} = 0, u(L,t) = 0, \frac{\partial u(L,t)}{\partial x} = 0, t\in[0,T].$$

Model 2. The left end is fixed and the right end is hinged:

$$u(0,t) = 0, \frac{\partial u(0,t)}{\partial x} = 0, u(L,t) = 0, \frac{\partial^2 u(L,t)}{\partial x^2} = 0, t\in[0,T].$$

Model 3. The left end is hinged and the right end is fixed:

$$u(0,t) = 0, \frac{\partial^2 u(0,t)}{\partial x^2} = 0, u(L,t) = 0, \frac{\partial u(L,t)}{\partial x} = 0, t\in[0,T].$$

Model 4. The ends of the beam are hinged:

$$u(0,t) = 0, \frac{\partial^2 u(0,t)}{\partial x^2} = 0, u(L,t) = 0, \frac{\partial^2 u(L,t)}{\partial x^2} = 0, t\in[0,T].$$

It is necessary to find an approximate and analytical solution to the above problems and to evaluate an approximate solution using an analytical solution. This estimate allows us to apply this approximate method to more complex Euler-Bernoulli beam problems. More precisely, to problems for which it is impossible to find an analytical solution.

Before the numerical solution of mechanical problems, equations, initial and boundary conditions are converted to dimensionless variables:

$$\xi = \frac{x}{L}, \tau = \frac{t}{t_0}, w(\xi,\tau) = \frac{u(x,t)}{u_0},$$

$t_0 = \sqrt{\frac{\rho AL^4}{EI}}$ is characteristic frequency,

$u_0 = \frac{PL^3}{EI}$ is characteristic scale of deflection.

Let us put all the variables in their places, get the equation in dimensionless variables:

$$\frac{\partial^4 w}{\partial \xi^4} + \frac{\partial^2 w}{\tau^2} + \lambda w = \cos\omega_0\tau\delta(\xi - v\tau), \xi \in [0,1], \tau \in [0, T_0]. \tag{4}$$

Here

$\omega_0 = \omega t_0$ is dimensionless frequency,

$\lambda = \frac{KL^4}{EI}$ is dimensionless Base factor,

$v = \frac{Vt_0}{L}$ is dimensionless speed,

$T_0 = \frac{T}{t_0}$ is the dimensionless duration of the period under consideration.

Dimensionless initial conditions:

$$w(\xi, 0) = 0, \frac{\partial w(\xi, 0)}{\partial \tau} = 0 = 0, \xi \in [0,1]. \tag{5}$$

Dimensionless boundary conditions:

Model 1:

$$w(0,\tau) = 0, \frac{\partial w(0,\tau)}{\partial \xi} = 0, w(1,\tau) = 0, \frac{\partial w(1,\tau)}{\partial \xi} = 0, \tau \in [0, T_0], \tag{6}$$

Model 2:

$$w(0,\tau) = 0, \frac{\partial^2 w(0,\tau)}{\partial \xi^2} = 0, w(1,\tau) = 0, \frac{\partial w(1,\tau)}{\partial \xi} = 0, \tau \in [0, T_0], \tag{7}$$

Model 3:

$$w(0,\tau) = 0, \frac{\partial w(0,\tau)}{\partial \xi} = 0, w(1,\tau) = 0, \frac{\partial^2 w(1,\tau)}{\partial \xi^2} = 0, \tau \in [0, T_0], \tag{8}$$

Model 4:

$$w(0,\tau) = 0, \frac{\partial^2 w(0,\tau)}{\partial \xi^2} = 0, w(1,\tau) = 0, \frac{\partial^2 w(1,\tau)}{\partial \xi^2} = 0, \tau \in [0, T_0], \tag{9}$$

Thus, we have four problems that need to be solved in an approximate and analytical ways.

3. Problem-solving methods

3.1. **Method of quadratures with derivatives.** To solve the above problems, we apply the following approach. Since the intensity of the action on the beam is a harmonic function, and the desired function that determines the forced oscillations of the same frequency ω_0 is represented as follows: $w(\xi, \tau) = y(\xi)\cos\omega_0\tau$, where $y(\xi)$ is the unknown function.

Since the function $\delta(\xi - v\tau)$ depends on time, it will be necessary to find $y(\xi)$ for each time layer. Therefore, moving to finite differences in time and substituting the above expressions into equation (4), we have a boundary value problem in ordinary differential equations:

$$y_j^{(4)}(\xi) = \mu y_j(\xi) + \delta(\xi - v\tau_j), \tag{10}$$

where $\mu = \omega_0^2 - \lambda$.

Boundary Conditions:

Model 1:

$$y_j(0) = \frac{dy_j}{dx}(0) = y_j(1) = \frac{dy_j}{dx}(1) = 0, j = \overline{0, M}, \tag{11}$$

Model 2:

$$y_j(0) = \frac{dy_j}{dx}(0) = y_j(1) = \frac{d^2y_j}{dx^2}(1) = 0, j = \overline{0, M}, \tag{12}$$

Model 3:

$$y_j(0) = \frac{d^2y_j}{dx^2}(0) = y_j(1) = \frac{dy_j}{dx}(1) = 0, j = \overline{0, M}, \tag{13}$$

Model 4:

$$y_j(0) = \frac{d^2y_j}{dx^2}(0) = y_j(1) = \frac{d^2y_j}{dx^2}(1) = 0, j = \overline{0,M}, \tag{14}$$

where $y_j(\xi)$ is the desired function is in the j time layer, $\tau_j = \Delta\tau j, j = \overline{0,M}, \Delta\tau$ is the step in time.

To solve boundary value problems (10) – (14), equation (10) is reduced to the Fredholm integral equation of the second kind [5]:

$$y_j(\xi) = \mu \int_0^1 G(\xi, s)y_j(s)ds + \int_0^1 G(\xi, s)\delta(s - v\tau_j)ds, \tag{15}$$

where $G(x, s)$ is the Green function. Under the conditions for determining the Green function, all boundary conditions (11) - (14) are taken into account. The second integral in equations (15)

$$\int_0^1 G(\xi, s)\delta(s - v\tau_j)ds = G(\xi, v\tau_j).$$

Hence, equation (15) will take the following form

$$y_j(\xi) = \mu \int_0^1 G(\xi, s)y_j(s)ds + G(\xi, v\tau_j).$$

To solve the integral equation (16), we apply the method of optimal quadrature formulas with derivatives in space $L_2^{(m)}(0, 1)$ [6]. In this method, the integral of (16) is replaced by the sum, i.e.

$$\int_0^1 G(\xi_i, s)y_j(s)ds \cong \sum_{k=0}^{m}\sum_{\beta=0}^{N} C_{i\beta}^{(k)} y_{\beta j}^k, i = \overline{0,N}, j = \overline{0,M}, \tag{17}$$

Here ξ_i is the nodes of the grid, $\xi_i = ih,\ i = \overline{0,N},\ N = h^{-1},\ h$ is the grid step, $y_{\beta j}^{(k)} = y_j^k(\xi_\beta),\ k = \overline{0,m-1},\ C_{i\beta}^{(k)}$ are the optimal coefficients of the quadrature formula (17).

The optimal coefficients of the quadrature formula (17) are determined using the following formulas:

$$C_{i0}^{(k)} = \frac{p_{ik}}{2} + h^{-1}(-1)^k[F_{ik1} - F_{ik0}],$$

$$C_{i\beta}^{(k)} = h^{-1}(-1)^k[F_{ik\beta-1} - 2F_{ik\beta} + F_{ik\beta+1}], \beta = \overline{1,N-1},$$

$$C_{iN}^{(k)} = \frac{p_{ik}}{2} + h^{-1}(-1)^k[F_{ikN-1} - F_{ikN}],$$

$$i = \overline{0,N}, k = \overline{0,m-1}, m = 1, 2, ...,$$

where

$$F_{ik\beta} = f_{ik\beta} - \sum_{\nu=0}^{k-1}\sum_{\gamma=0}^{N}(-1)^\nu C_{i\gamma}^{(\nu)} \frac{(h\beta - h\gamma)^{k-\nu+1} sign(h\beta - h\gamma)}{2(k-\nu+1)!},$$

$$f_{ik\beta} = (-1)^k \int_0^1 G^{(k)}(\xi_i, s)\frac{(s - h\beta)^{k+1} sign(s - h\beta)}{2(k+1)!}ds,$$

$$g_{ik} = \int_0^1 G^{(k)}(\xi_i, s)s^k ds, G^{(k)}(\xi_i, s) = \frac{\partial^k}{\partial\xi^k}G(\xi, s)|_{\xi=\xi_i},$$

$$p_{ik} = \frac{g_{ik}}{k!} - \sum_{\nu=0}^{k-1}\sum_{\gamma=0}^{N} C_{i\gamma}^{(k)} \frac{(h\gamma)^{k-\nu}}{(k-\nu)!},$$

$$i = \overline{0,N}, k = \overline{0,m-1}, m = 1, 2, ...$$

Applying the optimal quadrature formula (17) to the integral equation (16) and moving on to the difference equations, we get:

$$y_{ij} - \mu \sum_{k=0}^{m-1} \sum_{\beta=0}^{N} C_{i\beta}^{(k)} y_{\beta j}^{(k)} = G_{ij}, i = \overline{0,N}, j = \overline{0,M}, m = 1,2,.... \tag{18}$$

Where $G_{ij} = G(\xi_i,\ v\tau_j)$.

In the system of linear algebraic equations (18) at each time layer, the number of equations is $N+1$, and the number of unknowns $m(N+1)$. To obtain the remaining equations, we differentiate the integral equation (16) $m-1$ times with respect to the variable ξ:

$$\begin{gathered} y_j'(\xi) - \mu \int_0^1 G_\xi'(\xi,s) y_j(s) ds = G_\xi'(\xi, v\tau_j), \\ y_j''(\xi) - \mu \int_0^1 G_\xi''(\xi,s) y_j(s) ds = G_\xi''(\xi, v\tau_j), \\ \ldots\ \ldots\ \ldots \\ y_j^{(m-1)}(\xi) - \mu \int_0^1 G_\xi^{(m-1)}(\xi,s) y_j(s) ds = G_\xi^{(m-1)}(\xi, v\tau_j), \\ j = \overline{0,M}, m = 1,2,.... \end{gathered} \tag{19}$$

Furthermore, applying the optimal quadrature formula (17) for each equation (19) and moving on to the difference equations, we obtain a system of linear algebraic equations to determine the values of the desired function $y_j(\xi_i)$ and the values of the derivatives $y_j'(\xi_i)$, $y_j''(\xi_i)$, ..., $y_j^{(m-1)}(\xi_i)$ in the nodes of the grid $\xi_i (i = \overline{0,N})$:

$$\begin{gathered} y_{ij} - \mu \sum_{k=0}^{m-1} \sum_{\beta=0}^{N} C_{i\beta}^{(k)} y_{\beta j}^{(k)} = G_{ij}, \\ y_{ij}' - \mu \sum_{k=0}^{m-1} \sum_{\beta=0}^{N} C_{i\beta}^{(k)} y_{\beta j}^{(k)} = G_{ij}', \\ \ldots\ \ldots\ \ldots \\ y_{ij}^{(m-1)} - \mu \sum_{k=0}^{m-1} \sum_{\beta=0}^{N} C_{i\beta}^{(k)} y_{\beta j}^{(k)} = G_{ij}^{(m-1)}. \end{gathered} \tag{20}$$

From the solution on each time layer j of the linear system of algebraic equations (20), the first $N+1$ terms, i.e. $y_{ij} (i = \overline{0,N})$ will be the numerical solution of the integral equation (16).

Now let's move on to the function. Then $w_{ij} = w(\xi_i, \tau_j) = y_{ij} cos\omega_0\tau_j,\ i = \overline{0,N},\ j = \overline{0,M}$.

3.2. **Analytical solution.** To find an analytical solution to our problems, we define the Green function for a homogeneous equation [5]:

$$y^{(4)}(\xi) + \mu y(\xi) = 0, \xi \in [0,1]$$

taking into account the boundary conditions of each model (11) - (14). Each model will have its own Green function. After that, you can write an analytical solution to the inhomogeneous equation

$$y(\xi,\tau) = \int_0^1 G(\xi,s)\delta(s - v\tau) ds = G(\xi, v\tau).$$

Hence $w(\xi,\tau) = G(\xi, v\tau)\cos\omega_0\tau$.

Table 1. Initial information for beam deflection calculation

№	Parameter names	Unit	Parameter values
1	Coefficient of elasticity	N/m^2	2.0×10^{11}
2	Moment of inertia of the section	m^4	0.0001
3	Density of Beam Material	kg/m^3	7850
4	Cross-sectional Area of Beam	m^2	0.03
5	Elastic Base Bed Coefficient	N/m^2	1.0E+7
6	Beam Length	m	10
7	Beam Load	N	50000
8	Load velocity	m/s	50
9	Oscillation frequency	-	0
10	Duration of the study period	s	10

4. Numerical results

According to the above algorithm, a program is compiled in Matlab. Thanks to the Matlab language capability, the Green function is obtained in analytical form. All operations are performed using the program. For the numerical experiment, the initial information given in Table 1 is adopted.

For the integral equations and their right-hand sides in each model, the Green function is obtained:
Model 1:

$$G(x,\xi) = \begin{cases} -x^2(\xi-1)^2(x-3\xi+2x\xi)/6, & \text{if } x \le \xi, \\ -\xi^2(x-1)^2(\xi-3x+2x\xi)/6, & \text{if } x > \xi, \end{cases}$$

Model 2:

$$G(x,\xi) = \begin{cases} -x(\xi-1)(\xi^2-2\xi+x^2)/6, & \text{if } x \le \xi, \\ -\xi\,(x-1)\,(\xi^2-2x+x^2)/6, & \text{if } x > \xi, \end{cases}$$

Model 3:

$$G(x,\xi) = \begin{cases} x^2((-\xi^3+3\xi^2-2)x+(3\xi^3-9\xi^2+6\xi))/12, & \text{if } x \le \xi, \\ \xi^2(x-1)(2\xi-6x+2\xi x-\xi x^2+3x^2)/12, & \text{if } x > \xi, \end{cases}$$

Model 4:

$$G(x,\xi) = \begin{cases} -x\,(\xi-1)^2\,(\xi x^2-3\xi+2x^2)/12, & \text{if } x \le \xi, \\ -\xi\,(x-1)^2\,(\xi^2 x-3x+2\xi^2)/12, & \text{if } x > \xi, \end{cases}$$

Next, for the analytical solution, the Green function for each model is also obtained: Model 4:

$$G(x,\xi) = \begin{cases} G_1(x,\xi), & \text{if } x \le \xi, \\ G_2(x,\xi), & \text{if } x > \xi, \end{cases}$$

Model 1:

$$G_1(x,\xi) = (\sin(\beta^*\xi - \beta x) - \sinh(\beta^*\xi - \beta x) - \cos(\beta^*)\sinh(\beta^* - \beta^*\xi + \beta x) +$$

$$+\cosh(\beta^*)\sin(\beta^* - \beta^*\xi + \beta x) + \cos(\beta^*\xi)\sinh(\beta x) - \cosh(\beta^*\xi)\sin(\beta x) - \sin(\beta^*\xi)\cosh(\beta x) +$$

$$+\sinh(\beta^*\xi)\cos(\beta x) - \cos(\beta^*\xi - \beta^* + \beta x)\sinh(\beta^*) + \cosh(\beta^*\xi - \beta^* + \beta x)\sin(\beta^*) -$$

$$-\cos(\beta^* - \beta x)\sinh(\beta^*\xi - 1)) + \cosh(\beta^* - \beta x)\sin(\beta^*\xi - 1)) + \cos(\beta^*(\xi-1))\sinh(\beta^* - \beta x) -$$

$$-\cosh(\beta^*\xi - 1))\sin(\beta^* - \beta x))/(4{\beta^*}^3(\cos(\beta^*)\cosh(\beta^*) - 1));$$

$$G_2(x,\xi) = -(\sin(\beta^*\xi - \beta x) - \sinh(\beta^*\xi - \beta x) + \cos(\beta^*)\sinh(\beta^* + \beta^*\xi - \beta x) -$$

$$-\cosh(\beta^*)\sin(\beta^* + \beta^*\xi - \beta x) - \cos(\beta^*\xi)\sinh(\beta x) + \cosh(\beta^*\xi)\sin(\beta x) +$$

$$+\sin(\beta^*\xi)\cosh(\beta x) - \sinh(\beta^*\xi)\cos(\beta x) + \cos(\beta^*\xi - \beta^* + \beta x)\sinh(\beta^*) -$$

$$-\cosh(\beta^*\xi-\beta^*+\beta x)\sin(\beta^*)+\cos(\beta^*-\beta x)\sinh(\beta^*\xi-1))-\cosh(\beta^*-\beta x)\sin(\beta^*\xi-1))-$$
$$-\cos(\beta^*(\xi-1))\sinh(\beta^*-\beta x)+\cosh(\beta^*(\xi-1))sin(\beta^*-\beta x))/(4{\beta^*}^3(\cos(\beta^*)\cosh(\beta^*)-1));$$

Model 2:

$$G_1(x,\xi)=(\cos(\beta^*\xi)\sin(\beta x))/(2{\beta^*}^3)-(\cosh(\beta^*\xi)\sinh(\beta x))/(2{\beta^*}^3)-$$
$$-(\cot(\beta^*)\sin(\beta^*\xi)\sin(\beta x))/(2{\beta^*}^3)+(\cosh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x))/(2\sinh(\beta^*){\beta^*}^3);$$
$$G_2(x,\xi)=(\sin(\beta^*\xi)\cos(\beta x))/({\beta^*}^3)-(\sinh(\beta^*\xi)\cosh(\beta x))/(2{\beta^*}^3)-$$
$$-(\cot(\beta^*)\sin(\beta^*\xi)\sin(\beta x))/(2{\beta^*}^3)+(\cosh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x))/(2\sinh(\beta^*){\beta^*}^3);$$

Model 3:

$$G_1(x,\xi)=(\cosh(\beta^*)\cos(\beta^*-\beta^*\xi+\beta x)+\sinh(\beta^*)\sin(\beta^*-\beta^*\xi+\beta x)-\cos(\beta^*\xi-\beta^*+\beta x)\cosh(\beta^*)-$$
$$-\sin(\beta^*\xi-\beta^*+\beta x)\sinh(\beta^*)-2\cosh(\beta^*)\sinh(\beta x)\sin(\beta^*(\xi-1))+2\sinh(\beta^*)\cosh(\beta x)\sin(\beta^*(\xi-1))+$$
$$+2\cosh(\beta^*)\sinh(\beta^*\xi)\sin(\beta^*-\beta x)-2\sinh(\beta^*)\cosh(\beta^*\xi)\sin(\beta^*-\beta x)+2\cos(\beta^*)\cosh(\beta^*)\sinh(\beta^*\xi)$$
$$\sinh(\beta x)-2\cos(\beta^*)sinh(\beta^*)\cosh(\beta^*\xi)\sinh(\beta x)-2\cosh(\beta^*)\sin(\beta^*)\sinh(\beta^*\xi)\cosh(\beta x)+$$
$$+2\sin(\beta^*)\sinh(\beta^*)\cosh(\beta^*\xi)\cosh(\beta x))/(4{\beta^*}^3(\cos(\beta^*)\sinh(\beta^*)-\cosh(\beta^*)\sin(\beta^*)));$$
$$G_2(x,\xi)=(\cosh(\beta^*)\cos(\beta^*+\beta^*\xi-\beta x)+\sinh(\beta^*)\sin(\beta^*+\beta^*\xi-\beta x)-\cos(\beta^*\xi-\beta^*+$$
$$+\beta x)\cosh(\beta^*)-\sin(\beta^*\xi-\beta^*+\beta x)\sinh(\beta^*)-2\cosh(\beta^*)\sinh(\beta x)\sin(\beta^*(\xi-1))+$$
$$+2\sinh(\beta^*)\cosh(\beta x)\sin(\beta^*(\xi-1))+2\cosh(\beta^*)\sinh(\beta^*\xi)\sin(\beta^*-\beta x)-$$
$$-2\sinh(\beta^*)\cosh(\beta^*\xi)\sin(\beta^*-\beta x)+2\cos(\beta^*)\cosh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x)-$$
$$-2\cos(\beta^*)\sinh(\beta^*)\sinh(\beta^*\xi)\cosh(\beta x)-2\cosh(\beta^*)\sin(\beta^*)\cosh(\beta^*\xi)\sinh(\beta x)+$$
$$+2\sin(\beta^*)\sinh(\beta^*)\cosh(\beta^*\xi)\cosh(\beta x))/(4{\beta^*}^3(\cos(\beta^*)\sinh(\beta^*)-\cosh(\beta^*)\sin(\beta^*)));$$

Model 4:

$$G_1(x,\xi)=-(\sin(\beta^*\xi)\sinh(\beta x)+\sinh(\beta^*\xi)\sin(\beta x)-\cos(\beta^*)\cosh(\beta^*)\sin(\beta^*\xi)\sin(\beta x)-$$
$$-\cos(\beta^*)\sinh(\beta^*)\cos(\beta^*\xi)\sin(\beta x)+\cosh(\beta^*)\sin(\beta^*)\cos(\beta^*\xi)\sin(\beta x)-$$
$$-\cos(\beta^*)\cosh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x)+\cos(\beta^*)\sinh(\beta^*)\cosh(\beta^*\xi)\sinh(\beta x)-$$
$$-\cosh(\beta^*)\sin(\beta^*)\cosh(\beta^*\xi)\sinh(\beta x)-\sin(\beta^*)\sinh(\beta^*)\sin(\beta^*\xi)\sin(\beta x)+$$
$$+\sin(\beta^*)\sinh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x))/(2{\beta^*}^3(\cos(\beta^*)\sinh(\beta^*)-\cosh(\beta^*)\sin(\beta^*)));$$
$$G_2(x,\xi)=-(\sin(\beta^*\xi)\sinh(\beta x)+\sinh(\beta^*\xi)\sin(\beta x)-\cos(\beta^*)\cosh(\beta^*)\sin(\beta^*\xi)\sin(\beta x)-$$
$$-\cos(\beta^*)\sinh(\beta^*)\sin(\beta^*\xi)\cos(\beta x)+\cosh(\beta^*)\sin(\beta^*)\sin(\beta^*\xi)\cos(\beta x)-$$
$$-\cos(\beta^*)\cosh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x)+\cos(\beta^*)\sinh(\beta^*)\sinh(\beta^*\xi)\cosh(\beta x)-$$
$$-\cosh(\beta^*)\sin(\beta^*)\sinh(\beta^*\xi)\cosh(\beta x)-\sin(\beta^*)\sinh(\beta^*)\sin(\beta^*\xi)\sin(\beta x)+$$
$$+\sin(\beta^*)\sinh(\beta^*)\sinh(\beta^*\xi)\sinh(\beta x))/(2{\beta^*}^3(\cos(\beta^*)\sinh(\beta^*)-\cosh(\beta^*)\sin(\beta^*))),$$

where β is the solution of the characteristic equation $\beta^4+\mu=0$, β^*- is a conjugate complex number.

On the basis of these calculation formulas, deflection calculations were carried out for all models approximately by the method of quadrature formulas with derivatives and the values of the analytical solution. The results of the calculations are given in Table 2 and Table 3.

Table 2. Initial information for beam deflection calculation

Time	Model 1	Model 2	Model 3	Model 4
$0,0$	$0.000E+00$	$0.000E-00$	$0.000E+00$	$0.000E+00$
$0,1$	$5.278E-08$	$1.330E-07$	$3.086E-08$	$1.072E-07$
$0,2$	$4.088E-07$	$3.065E-07$	$1.886E-07$	$3.150E-07$
0.3	$3.071E-07$	$2.965E-07$	$1.638E-07$	$2.626E-07$
0.4	$7.154E-07$	$3.002E-07$	$3.248E-07$	$3.534E-07$
0.5	$5.271E-07$	$2.638E-07$	$2.669E-07$	$2.904E-07$
0.6	$6.925E-07$	$2.139E-07$	$3.609E-07$	$2.249E-07$
0.7	$5.013E-07$	$1.798E-07$	$2.955E-07$	$1.792E-07$
0.8	$3.088E-07$	$2.567E-07$	$2.438E-07$	$2.158E-07$
0.9	$2.172E-07$	$2.453E-07$	$1.978E-07$	$1.908E-07$
1.0	$2.685E-07$	$2.516E-07$	$2.162E-07$	$2.383E-07$
Average	$3.636E-07$	$2.225E-07$	$2.081E-07$	$2.161E-07$

Table 3. Average values of absolute error for models corresponding to the method parameter m.

m	Model 1	Model 2	Model 3	Model 4
1	9.263E-05	2.125E-05	1.990E-05	2.869E-05
2	2.235E-06	4.894E-07	4.959E-07	9.053E-07
3	1.142E-06	2.997E-07	3.132E-07	4.674E-07
4	3.636E-07	2.081E-07	2.161E-07	2.225E-07
5	3.636E-07	2.081E-07	2.161E-07	2.225E-07
6	3.636E-07	2.081E-07	2.161E-07	2.225E-07

5. Conclusion

Based on the results obtained, the following conclusions can be drawn:

- In the calculation process, the optimal value of the parameter of the method of derivative quadrature formulas was determined ($m = 4$);
- An analytical solution has been obtained using the Green function;
- Green's function is obtained in symbolic form, which is used to solve the integral equation;
- a high order of accuracy of the approximate solution was obtained, which is proved by the values of absolute error;
- the possibility of considering another model, i.e. other boundary conditions must be taken into account in the construction of the Green function.

The proposed method can be generalized to solve more complex problems of Euler-Bernoulli beams.

References

[1] Timoshenko S. P., Method of analysis of statical and dynamical stresses in rail. Proceedings of the 2nd International Congress for Applied Mechanics, Zurich, (1927), P. 407–418.

[2] Vakulenko S. P., Volosov K. A., Volosova N. K., On the method of assessing the state of the railway track. World of Transport, (2016) Vol. 14, Iss. 3, P. 20–35.

[3] Kurbatsky E. N., Zernov I. I., Badina E. S., Application of generalized functions and integral Fourier transform in modeling the effect of a moving load on a beam lying on an elastic base. Transport Structures, (2023) Vol. 10, Iss. 3,

[4] Demchenko D. B., Mayatsky I. A., Polismakov A. I., Numerical Realization of the Problem of Bending a Beam-Strips on an Elastic Base by the Finite Difference Method. Young Researcher of the Don, (2017), Iss. 2, P. 81–94.

[5] Kalashnikov A. L., Methods of Approximate Solution of Integral Equations of the Second Kind. Nizhny Novgorod State University named after N. I. Lobachevsky, Nizhny Novgorod, (2017).

[6] Shadimetov K. M., Usmanov K. I., Approximate solution of linear Fredholm integral equations of the second kind by the method of optimal quadratures. Problems of Computational and Applied Mathematics, (2023), Iss. 5, P. 1–23.

Usmanov Kh.I.,
V.I.Romanovskiy Institute of Mathematics,
Uzbekistan Academy of Sciences,
Tashkent, Uzbekistan;
e-mail: hojiakbar170853@mail.ru

Uzbek Mathematical Journal
2026, Volume 70, Issue 2, pp. 208–217
DOI: 10.29229/uzmj.2026-2-24

Geometric and topological properties of the Minkowski operator on convex sets

Zaitov A., Nuritdinov J.

Abstract. This paper systematically establishes the fundamental geometrical and topological properties of the Minkowski operator on convex sets in an Euclidean space. The convexity, closedness, and boundary properties of the Minkowski sum and difference are analyzed. In addition, results concerning the Minkowski sum and difference of open and closed sets are presented. The paper also constructs illustrative examples to clarify each of the obtained results.

Keywords: Convex sets, Minkowski difference, Minkowski sum, homothetic set, boundary of a set, compact set
MSC (2020): 52A20, 54C10

1. Preliminaries

Definition 1.1. [1] . Let X and Y be non-empty sets in $\mathbb{R}^n$. The geometric (or Minkowski) sum of X and Y is defined as

$$X + Y = \{x + y : x \in X, y \in Y\}$$

and the geometric (or Minkowski) difference is defined as

$$X \dot{-} Y = \{z \in \mathbb{R}^n : z + Y \subset X\}.$$

For a set $X \subset \mathbb{R}^n$ and a scalar $\lambda \in \mathbb{R}$, the scalar multiplication of X by λ is defined as

$$\lambda X = \{\lambda x : x \in X\}.$$

For any vector $a \in \mathbb{R}^n$ and non-empty set $X \subset \mathbb{R}^n$, the Minkowski sum of the vector a and the set X is defined as the Minkowski sum of the singleton a and X:

$$a + X = \{a\} + X = \{a + x : x \in X\}.$$

Definition 1.2. [2] Let X be a nonempty set in $\mathbb{R}^n$ and let $\lambda > 0$ be a real number. The set $X_{(\lambda,t)}$ defined by $X_{(\lambda,t)} = \lambda X + t$ is called a homothetic set of X. Here $t \in \mathbb{R}^n$.

Let N be a compact set and let M a convex compact set in $\mathbb{R}^n$. Assume that a foliation $\mathcal{F} = \{L_\alpha : L_\alpha = \partial M_\alpha,\ \alpha \in A\}$ is given, consisting of the boundaries of homothetic subsets of M.

Definition 1.3. If there exist an index $\alpha \in A$ and an element $z \in \mathbb{R}^n$ such that the relation $z+N \subset M_\alpha$ holds, then the set N is said can embed into the foliation $\mathcal{F}$.

Definition 1.4. If for some index $\alpha_0 \in A$ the relation $z + N \subset M_{\alpha_0}$ holds, but for every index $\alpha \in A$ with $\alpha < \alpha_0$ the inclusion $z + N \subset M_\alpha$ does not hold then the set N is said can embed into the set M_{α_0}.

From this definition it is easy to see that if a compact set N can embed into the set M_{α_0}, then the dimension of the geometric difference $M_{\alpha_0} \dot{-} N$ is smaller than the dimension of the space $\mathbb{R}^n$.

Definition 1.5. If for some index $\alpha_0 \in A$ the Minkowski difference $M_{\alpha_0} \dot{-} N$ consists of a single point, that is, $M_{\alpha_0} \dot{-} N = \{a\}$ for some $a \in \mathbb{R}^n$, then the compact set N is said to be fixedly embedded into the set M_{α_0}.

2. Results on the Minkowski Difference of Convex Sets

Lemma 2.1. *For a convex set M in $\mathbb{R}^n$ and its homothetic set $M_\alpha = \alpha M$ (where $0 < \alpha \le 1$), equality $D = M \dot{-} M_\alpha$ implies equality $M = D + M_\alpha$.*

Proof: The equality $D = M \dot{-} M_\alpha$ and the definition of the Minkowski difference imply the inclusion

$$D + M_\alpha \subset M. \tag{2.1}$$

Let us show the inverse inclusion. For each vector $x \in M$ and number α $(0 < \alpha \le 1)$, denote the difference $x - \alpha x$ by d, i. e., $d = x - \alpha x = (1 - \alpha)x$. Since M is a convex set, for any point $y \in M$, we have

$$d + \alpha y = (1 - \alpha)x + \alpha y \in M.$$

Hence, $d + M_\alpha \subset M$, which means $d \in D$. Thus, for every point $x \in M$, there exists a point $d \in D$ such that the equality $x = \alpha x + d$ holds. This implies

$$M \subset D + M_\alpha \tag{2.2}$$

The relations (2.1) and (2.2) together complete the proof of the lemma. □

Theorem 2.2. *Let M and N be nonempty convex sets in $\mathbb{R}^n$, and suppose that $M_\alpha \dot{-} N \ne \varnothing$ for some $\alpha \in (0, 1]$. Then the equality*

$$M \dot{-} N = (M \dot{-} M_\alpha) + (M_\alpha \dot{-} N) \tag{2.3}$$

holds.

Proof: Let $z \in M \dot{-} N$. It must be shown that there exist points $z_1 \in M \dot{-} M_\alpha$ and $z_2 \in M_\alpha \dot{-} N$ satisfying the equality $z = z_1 + z_2$.

For an element $z \in M \dot{-} N$, using the definition of the Minkowski difference, we can write the relation $z + N \subset M$. Hence, for every $c \in N$, there exists an element $a \in M$ such that the equality $z + c = a$ holds.

Since $M_\alpha = \alpha M$, by Lemma 2.1, for each $a \in M$ there exist points $z_1 \in M \dot{-} M_\alpha$ and $m_\alpha \in M_\alpha$ such that a can be expressed in the form $a = m_\alpha + z_1$. Taking into account the relation $z + N \subset M$, for every $c \in N$ we have $z + c = m_\alpha + z_1$. From this equality we can write $z - z_1 + c = m_\alpha$. This implies that $z - z_1 \in M_\alpha \dot{-} N$. Thus, for any element $z \in M \dot{-} N$, we can always find a $z_1 \in M \dot{-} M_\alpha$ such that $z_2 = z - z_1 \in M_\alpha \dot{-} N$.

Now, let $z \in (M \dot{-} M_\alpha) + (M_\alpha \dot{-} N)$. Then there exist elements $z_1 \in M \dot{-} M_\alpha$ and $z_2 \in M_\alpha \dot{-} N$ such that $z_1 + z_2 = z$. From the relation $z_1 \in M \dot{-} M_\alpha$, by the definition of the Minkowski difference, we can write $z_1 + M_\alpha \subset M$. Similarly, from $z_2 \in M_\alpha \dot{-} N$ we obtain $z_2 + N \subset M_\alpha$. From these two inclusions it follows that $z_1 + z_2 + N \subset M$, and hence $z_1 + z_2 \subset M \dot{-} N$. Theorem 2.2 is proved. □

Let us consider an ordered set A such that for every pair $\alpha, \beta \in A$, $\alpha \ne \beta$, there exists an element $\gamma \in A$ satisfying $\alpha < \gamma < \beta$. As an example, the interval $[0, 1]$ can be taken as the set A.

Suppose a convex compact set M_1 in $\mathbb{R}^n$ and one of its interior points $a \in \operatorname{int} M_1$ are given. Denote its boundary by $\partial M_1 = L_1$. Convex subsets M_α of M_1 are chosen such that their boundaries $\partial M_\alpha = L_\alpha$, $\alpha \in A$, satisfy the following conditions:

1) for all $\alpha, \beta \in A$, $\alpha \ne \beta$, we have $L_\alpha \cap L_\beta = \varnothing$;

2) $\bigcup_{\alpha \in A} L_\alpha = M_1 \setminus \{a\}$;

3) for any $\alpha, \beta \in A$ with $\alpha \le \beta$, the condition $M_\beta \dot{-} M_\alpha \ne \varnothing$ holds.

The family $\mathcal{F} = \{L_\alpha \colon \alpha \in A\}$ is called a foliation, and each L_α is called a leaf, $\alpha \in A$.

Now consider an arbitrary convex compact set M in $\mathbb{R}^n$. Choose a point $a \in \operatorname{int} M$. For each number $\alpha \in (0, 1]$, the set $\alpha M + (1 - \alpha)a$ is a convex compact subset of M. Introducing the notation $L_\alpha = \partial(\alpha M + (1 - \alpha)a)$, we form the family $\mathcal{F} = \{L_\alpha \colon 0 < \alpha \le 1\}$. This family consists of the boundaries of subsets of M that are homothetic to M. At $\alpha = 0$, a special case arises: $M_0 = \{a\}$. This set is generally not homothetic to the given set $M = M_1$.

For the constructed family $\mathcal{F} = \{L_\alpha \colon 0 < \alpha \le 1\}$, the following statement follows directly from the definition of the Minkowski difference.

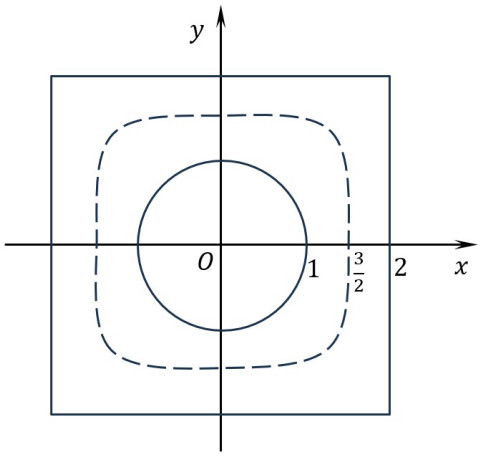


FIGURE 1.

Lemma 2.3. *The family $\mathcal{F}$ of boundaries of homothetic subsets of a convex compact set M in $\mathbb{R}^n$ satisfies the three conditions in the definition of a foliation, namely:*

1) for all $\alpha, \beta \in A$, $\alpha \neq \beta$, we have $L_\alpha \cap L_\beta = \varnothing$;

2) $\bigcup_{\alpha \in A} L_\alpha = M \setminus \{a\}$;

3) for any $\alpha, \beta \in A$ with $\alpha \leq \beta$, the condition $M_\beta \dot{-} M_\alpha \neq \varnothing$ holds.

It should be noted that for any family of subsets of M satisfying conditions 1), 2), and 3), Lemma 2.1 and Theorem 2.2 may not necessarily hold.

Example 2.4. Consider convex sets in the plane $\mathbb{R}^2$ with boundaries defined as following:

$$\partial M_1 = \{(x,\, y)\colon\ x = \pm 2 \ \text{ and } \ -2 \leq y \leq 2\} \cup \{(x,\, y)\colon\ -2 \leq x \leq 2 \ \text{ and } \ y = \pm 2\};$$

$$\partial M_{\frac{1}{2}} = \{(x,\, y)\colon\ x^2 + y^2 = 1\};$$

$$\partial M_t = \begin{cases} \left\{2(1-t)X + (2t-1)Y : X \in \partial M_{\frac{1}{2}}, Y \in \partial M_1\right\}, & \text{if } \frac{1}{2} \leq t \leq 1, \\ \left\{2tX\colon\ X \in \partial M_{\frac{1}{2}}\right\}, & \text{if } 0 < t \leq \frac{1}{2}. \end{cases}$$

Take the origin as the fixed interior point of M_1, i. e., let $a = (0,\, 0)$ (see Fig. 1).

For the family of sets M_t, $t \in (0, 1]$, the family $\{\partial M_t\colon\ t \in (0, 1]\}$ satisfies conditions 1), 2), and 3) stated above. However, the conclusion of Lemma 2.1 does not hold.

Indeed, consider the square $M_1 = \{(x,\, y)\colon\ -2 \leq x \leq 2,\ -2 \leq y \leq 2\}$ and the disk $M_{\frac{1}{2}} = \{(x,\, y)\colon\ x^2 + y^2 \leq 1\}$. The Minkowski difference $D = M_1 \dot{-} M_{\frac{1}{2}}$ is the square with side length 1:

$$D = M_1 \dot{-} M_{\frac{1}{2}} = \{(x,\, y)\colon\ -1 \leq y \leq 1,\ -1 \leq x \leq 1\} \qquad \text{(see Fig. 2)}.$$

But the Minkowski sum of the square D and the disk $M_{\frac{1}{2}}$ is not equal to the set M_1 (Fig. 3) [3].

According to [1], for these sets the strict inclusion

$$\left(M_1 \dot{-} M_{\frac{1}{2}}\right) + M_{\frac{1}{2}} \subset M_1$$

holds. Consequently, the conclusion of Theorem 2.2 is also not satisfied for this family.

The following result is established by directly verifying the definitions of the Minkowski difference and sum, as well as by applying Theorem 2.2.

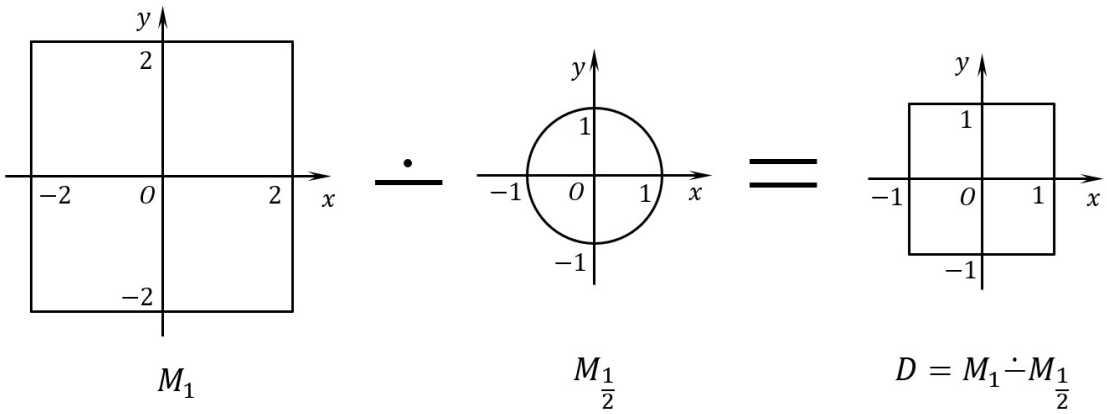


FIGURE 2.

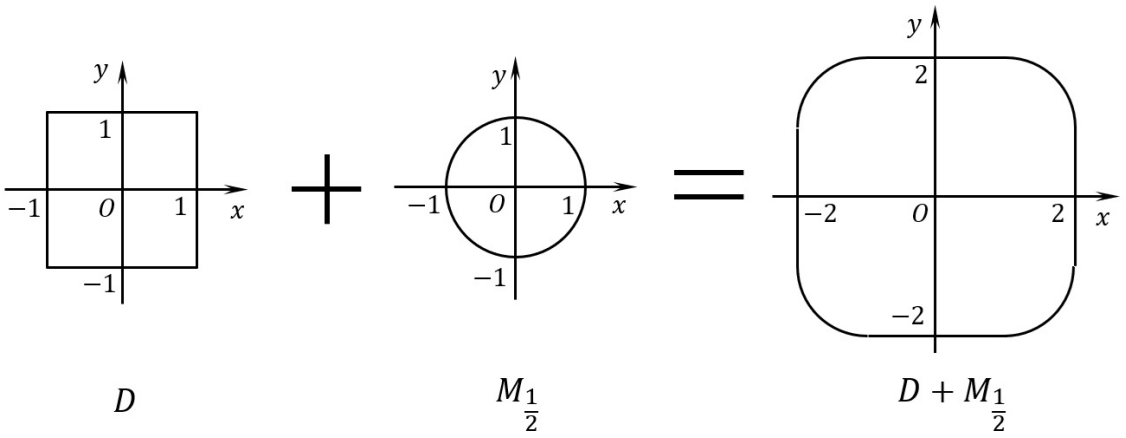


FIGURE 3.

Theorem 2.5. *If a compact set N is fixedly embedded in some set M_{α_0}, $\alpha \in (0, 1]$, then the equality*

$$M \dot{-} N = (M \dot{-} M_{\alpha_0}) + a \tag{2.4}$$

holds, where $M_{\alpha_0} \dot{-} N = \{a\}$.

The essence of this theorem is that in space $\mathbb{R}^n$, the problem of finding the Minkowski difference of an arbitrary compact set N from a convex compact set M can be reduced to the following, relatively simpler, problem: find the Minkowski difference from M of a set M_{α_0} that is homothetic to M and into which the compact set N is embedded (if it exists, fixedly embedded).

Problems of finding the Minkowski difference for mutually homothetic sets have been described in works [4], [5], [6].

The methods and examples presented above show that performing Minkowski operations in spaces of smaller dimension (for example, in $\mathbb{R}^2$) is easier and more efficient. However, in solving practical problems, one often has to work with sets given in higher-dimensional spaces. Under such conditions, it is convenient to decompose the given set defined by the Cartesian product into lower-dimensional sets and work with them.

Theorem 2.6. *For nonempty sets A, B, C and D in $\mathbb{R}^n$, the following equality holds:*

$$(A \times B) + (C \times D) = (A + C) \times (B + D). \tag{2.5}$$

Here $A \times B$ denotes the Cartesian product of sets.

Proof: For any point $x \in (A \times B) + (C \times D)$, by the definition of the Minkowski sum there exist points $x_1 \in A \times B$ and $x_2 \in C \times D$ such that

$$x = x_1 + x_2. \tag{2.6}$$

Since $x_1 \in A \times B$, by the definition of the Cartesian product it can be written as $x_1 = (a, b)$, where $a \in A$ and $b \in B$. Similarly, since $x_2 \in C \times D$, we have $x_2 = (c, d)$, where $c \in C$ and $d \in D$. Then equality (2.6) can be written in the form

$$x = x_1 + x_2 = (a, b) + (c, d).$$

By the definition of the Minkowski sum,

$$(a, b) + (c, d) = (a + c, b + d) \in (A + C) \times (B + D).$$

Now take an arbitrary element $z \in (A + C) \times (B + D)$. By the definition of the Cartesian product of sets, it can be written as an ordered pair

$$z = (z_1, z_2), \qquad z_1 \in (A + C), \ z_2 \in (B + D). \tag{2.7}$$

Since $z_1 \in (A + C)$, by the definition of the Minkowski sum there exist $a \in A$ and $c \in C$ such that $z_1 = a + c$. Similarly, from $z_2 \in (B + D)$ there exist $b \in B$ and $d \in D$ such that $z_2 = b + d$. Then expression (2.7) can be written as

$$z = (z_1, z_2) = (a + c, b + d) = (a, b) + (c, d). \tag{2.8}$$

Since $(a, b) \in A \times B$ and $(c, d) \in C \times D$, it follows that $z \in (A \times B) + (C \times D)$. Thus, Theorem 2.6 is proved. □

Lemma 2.7. *Let X and Y be nonempty sets, and let Z be a set homothetic to $X+Y$. If the condition $Z \subset X+Y$ is satisfied, then there exist nonempty sets X' and Y', homothetic to X and Y, respectively, such that $X' + Y' = Z$.*

Proof: Assume that $Z = \alpha(X + Y) + t$, where $t \in \mathbb{R}^n$ and $0 < \alpha \leq 1$. Then, for the sets $X' = \alpha X + t_X$ and $Y' = \alpha Y + t_Y$, the equality $X' + Y' = Z$ holds, where t_X and t_Y are any vectors with $t_X + t_Y = t$. Lemma 2.7 is proved. □

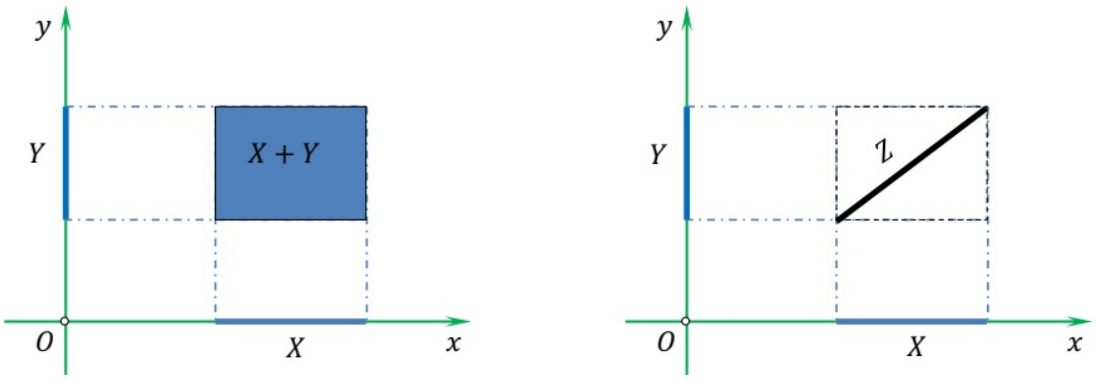


FIGURE 4.

The condition that the set Z be homothetic to the sum $X + Y$ is essential. Indeed, the set $Z = \left\{(x, y)\colon\ a \leq x \leq b;\ y = \frac{d-c}{b-a}(x - a) + c\right\}$ lies in the sum $X + Y = [a, b] \times [c, d]$ of the sets $X = [a, b] \times \{0\}$ and $Y = \{0\} \times [c, d]$. However, it is clear that the shape of the set Z is not similar to the shape of the sum $X + Y$; that is, Z is not a homothetic set. Moreover, there is no pair of sets $X' \subset X$ and $Y' \subset Y$ such that $X' + Y' = Z$ (see Fig. 4).

3. Topological Properties of Minkowski Operations

The density of topological spaces is one of the important concepts. The following statements can be established by direct verification.

Proposition 3.1. *If a set A is dense in a set $X \subset \mathbb{R}^n$, then for any vector $a \in \mathbb{R}^n$ the set $a + A$ is dense in $a + X$, and the set $A \dot{-} a$ is dense in $X \dot{-} a$.*

This statement can be generalized as follows.

Proposition 3.2. *If A is dense in X and B is dense in Y, then the set $A + B$ is dense in $X + Y$.*

To illustrate this statement consider the set $A = [2, 5]_{\mathbb{Q}}$ of all rational numbers in the interval $X = [2, 5]$, and the set $B = [8, 9]_{\mathbb{I}}$ of all irrational numbers in the interval $Y = [8, 9]$. Then the sum $A + B = [10, 14]_{\mathbb{I}}$ is the set of all irrational numbers in the interval $X + Y = [10, 14]$.

The following lemmas are used to represent a given set in the form of the Minkowski sum of two sets and make it possible to compute the Minkowski sum by decomposing it into parts.

Lemma 3.3. *Let X, Y and Z be nonempty sets such that $Z \cap (X + Y) \neq \varnothing$ and suppose that the set Z is homothetic to the sum $X + Y$. Then there exist sets U and V with $U \cap X \neq \varnothing$ and $V \cap Y \neq \varnothing$ such that $U + V = Z$.*

Proof: Denote the intersection $Z \cap (X + Y)$ by E, that is, $E = Z \cap (X + Y)$. Since for some $\alpha > 0$ and $t \in \mathbb{R}^n$ we have $E = \alpha(X + Y) + t$, by Lemma 2.7 there exist nonempty sets $X' = \alpha X + t_X$ and $Y' = \alpha Y + t_Y$ such that $X' + Y' = E$. Because $X' + Y' = E \subset Z$, by the definition of the Minkowski sum there exist sets $U \subset X'$ and $V \subset Y'$ satisfying $U + V = Z$. Consequently, $X' \cap X \neq \varnothing$ and $Y' \cap Y \neq \varnothing$. Thus, Lemma 3.3 is proved. □

From Figure 4 it can be seen that the condition that the set Z be homothetic to the sum $X + Y$ is essential.

Theorem 3.4. *For any nonempty sets X and Y, the inclusion*

$$\operatorname{int} X + \operatorname{int} Y \subset \operatorname{int}(X + Y)$$

holds. If, in addition, the sets X and Y are convex and $\dim X = \dim Y = \dim(X + Y)$, then the converse inclusion

$$\operatorname{int} X + \operatorname{int} Y \supset \operatorname{int}(X + Y)$$

also holds. Consequently, for such sets we have

$$\operatorname{int} X + \operatorname{int} Y = \operatorname{int}(X + Y).$$

Proof: Let $z \in \operatorname{int} X + \operatorname{int} Y$. Then there exist points $x \in \operatorname{int} X$ and $y \in \operatorname{int} Y$ such that $z = x + y$. Consider neighborhoods $B_x \subset X$ and $B_y \subset Y$. Then $B_x + B_y \subset X + Y$. Since B_x and B_y are open balls centered at x and y, respectively, their sum is an open set U. From the inclusion $U \subset X + Y$ it follows that $z \in \operatorname{int}(X + Y)$.

In general, the inclusion $\operatorname{int}(X + Y) \subset \operatorname{int} X + \operatorname{int} Y$ does not always hold. For example, if $\dim X < \dim(X + Y)$, then $\operatorname{int} X = \varnothing$, or if $\dim Y < \dim(X + Y)$, then $\operatorname{int} Y = \varnothing$.

Now assume that $\dim X = \dim Y = \dim(X + Y)$. Then

$$\operatorname{int} X + Y = X + \operatorname{int} Y = \operatorname{int} X + \operatorname{int} Y = \operatorname{int}(X + Y).$$

Theorem 3.4 is proved. □

We note that the convexity of the sets in the second part of Theorem 3.4 is essential.

Corollary 3.5. *a) for any convex sets X and Y in $\mathbb{R}^n$ with $\dim X = \dim Y = n$, the equality*

$$\operatorname{int} X + \operatorname{int} Y = \operatorname{int}(X + Y)$$

holds.

b) in the family $c(\mathbb{R}^n)$ of n-dimensional convex sets in $\mathbb{R}^n$, the Minkowski sum

$$+\colon c(\mathbb{R}^n) \times c(\mathbb{R}^n) \to c(\mathbb{R}^n)$$

is a continuous mapping.

Theorem 3.6. *Let X and Y be nonempty convex sets satisfying* $\dim X = \dim Y = \dim(X \dot{-} Y)$. *Then the inclusion* $\operatorname{int}(X \dot{-} Y) \subset \operatorname{int} X \dot{-} \operatorname{int} Y$ *holds.*

Proof: Take an arbitrary point $z \in \operatorname{int}(X \dot{-} Y)$. Then there exists an open neighborhood $B_r(z) = x \in \mathbb{R}^n\colon\ |z - x| < r$ such that $B_r(z) \subset X \dot{-} Y$. From here we obtain $B_r(z) + Y \subset X$. Since $\operatorname{int} Y \subset Y$, it follows that $B_r(z) + \operatorname{int} Y \subset B_r(z) + Y \subset X$. Because both $B_r(z)$ and $\operatorname{int} Y$ are open sets, the set $B_r(z) + \operatorname{int} Y$ is open. Hence, $B_r(z) + \operatorname{int} Y \subset \operatorname{int} X$. Therefore, $B_r(z) \subset \operatorname{int} X \dot{-} \operatorname{int} Y$, and consequently $z \in \operatorname{int} X \dot{-} \operatorname{int} Y$. Theorem 3.6 is proved. □

The existence of a point $z \in \operatorname{int} X \dot{-} \operatorname{int} Y$ that does not belong to $\operatorname{int}(X \dot{-} Y)$ can be seen from the following example. Let $X = [1, 5]$ and $Y = [3, 4]$. Then $\operatorname{int} X \dot{-} \operatorname{int} Y = [-2, 1]$, $\operatorname{int}(X \dot{-} Y) = (-2, 1)$. The points $z = -2$ and $z = 1$ have the stated property.

Lemma 3.7. *For arbitrary convex sets X and Y, the inclusion*

$$\mathbb{R}^n \setminus (X + Y) \subset (\mathbb{R}^n \setminus X) + (\mathbb{R}^n \setminus Y)$$

holds.

Proof: Let $z \in \mathbb{R}^n \setminus (X + Y)$. Then $z \notin X + Y$. This means that for any pair of vectors $x \in X$ and $y \in Y$ the equality $z = x + y$ does not hold. For each point $z \in \mathbb{R}^n \setminus (X + Y)$ and fixed points $x \in X$ and $y \in Y$ let us introduce the notations

$$Y_x(z) = \{y \in \mathbb{R}^n\colon\ z = x + y\}, \qquad X_y(z) = \{x \in \mathbb{R}^n\colon\ z = x + y\}.$$

Then, for every $z \in \mathbb{R}^n \setminus (X + Y)$, we obtain

$$\bigcup_{x \in X} Y_x(z) \subset (\mathbb{R}^n \setminus Y), \qquad \bigcup_{y \in Y} X_y(z) \subset (\mathbb{R}^n \setminus X).$$

On the other hand, by the definitions of the sets $Y_x(z)$ and $X_y(z)$ one has

$$z \in X_y(z) + Y_x(z) \subset \bigcup_{y \in Y} X_y(z) + \bigcup_{x \in X} Y_x(z) \subset (\mathbb{R}^n \setminus X) + (\mathbb{R}^n \setminus Y).$$

Thus, $\mathbb{R}^n \setminus (X + Y) \subset (\mathbb{R}^n \setminus X) + (\mathbb{R}^n \setminus Y)$.

Now we show that, in general, $\mathbb{R}^n \setminus (X + Y) \supset (\mathbb{R}^n \setminus X) + (\mathbb{R}^n \setminus Y)$ does not hold. For simplicity, it suffices to consider the case $n = 2$. Consider the following convex sets:

$$X = \{(x, y) \in \mathbb{R}^2\colon\ 1 \le x \le 4,\ 3 \le y \le 6\},$$

$$Y = \{(x, y) \in \mathbb{R}^2\colon\ 5 \le x \le 6,\ 1 \le y \le 2\}.$$

Then their Minkowski sum is the convex set

$$X + Y = \{(x, y) \in \mathbb{R}^2\colon\ 6 \le x \le 10,\ 4 \le y \le 8\}.$$

Consider also the sets

$$T = \{(x, y) \in \mathbb{R}^2\colon\ 8 \le x \le 10,\ 9 \le y \le 11\},$$

$$S = \{(x, y) \in \mathbb{R}^2\colon\ -2 \le x \le 0,\ -5 \le y \le -3\}.$$

Clearly, $T + S = X + Y$. Moreover, $T \subset \mathbb{R}^2 \setminus X$ and $S \subset \mathbb{R}^2 \setminus Y$. Hence,

$$X + Y \subset (\mathbb{R}^2 \setminus X) + (\mathbb{R}^2 \setminus Y).$$

Above, the inclusion $\mathbb{R}^2 \setminus (X + Y) \subset (\mathbb{R}^2 \setminus X) + (\mathbb{R}^2 \setminus Y)$ was established. From the last two inclusions we obtain

$$\mathbb{R}^2 = (X + Y) \cup \big(\mathbb{R}^2 \setminus (X + Y)\big) \subset (\mathbb{R}^2 \setminus X) + (\mathbb{R}^2 \setminus Y),$$

that is, $\mathbb{R}^2 = (\mathbb{R}^2 \setminus X) + (\mathbb{R}^2 \setminus Y)$. Lemma 3.7 is proved. □

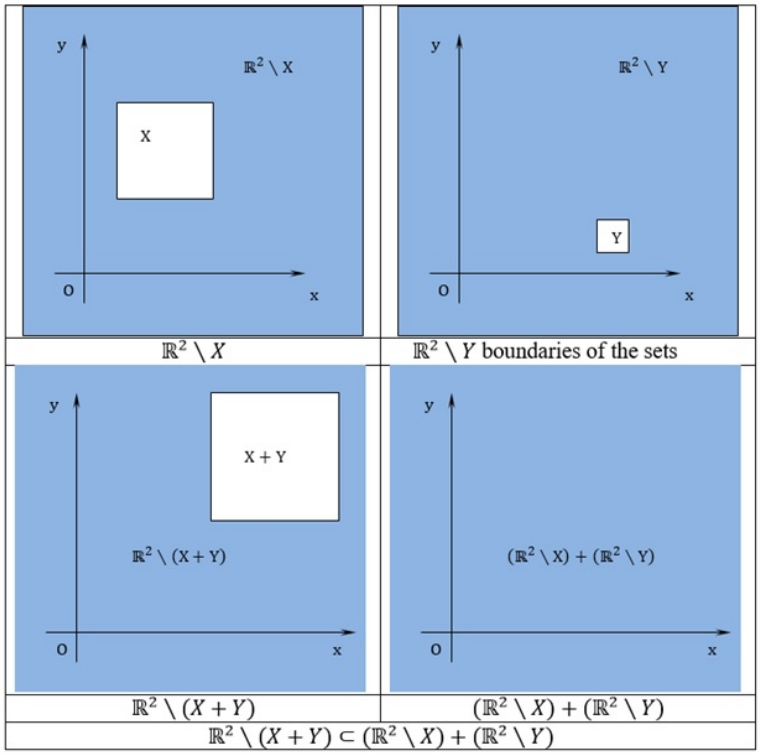


FIGURE 5.

An illustration of the proven lemma is presented in Figure 5.

Theorem 3.8. *For nonempty convex sets X, Y the following inclusions hold:*

$$\partial(X + Y) \subset \partial X + \partial Y \subset [X + Y].$$

Proof: Since the second inclusion in the conclusion of the theorem holds automatically, it suffices to show that any point $z \in \partial(X+Y)$ belongs to the set $\partial X+\partial Y$. By the definition of a boundary point, for any neighborhood $B_r(z)$ of the point z the relations $B_r(z)\cap(X+Y) \neq \varnothing$ and $B_r(z)\cap(\mathbb{R}^n\setminus(X+Y)) \neq \varnothing$ hold.

We can find x, $y \in \mathbb{R}^n$ such that for x and y we claim following

1) $z = x + y$,

2) $x \in \partial X$ and $y \in \partial Y$.

Suppose that for all x, $y \in \mathbb{R}^n$ satisfying condition 1), condition 2) does not hold. Then, for all x, $y \in \mathbb{R}^n$ such that $z = x + y$, it must be that either $x \notin \partial X$ and $y \in \partial Y$, or $x \in \partial X$ and $y \notin \partial Y$, or $x \notin \partial X$ and $y \notin \partial Y$. For clarity, let us assume $x \notin \partial X$ and $y \in \partial Y$. Then, either $x \in \operatorname{int} X$ or $x \in \mathbb{R}^n \setminus [X]$. In the first case, there exists an open ball $B_\varepsilon(x)$ such that $B_\varepsilon(x) \subset X$; in the second case, there exists an open ball $B_\varepsilon(x)$ such that $B_\varepsilon(x) \cap X = \emptyset$. In the first case,

$$B_\varepsilon(x + y) = B_\varepsilon(z) = B_\varepsilon(x) + y \subset X + Y,$$

which implies $z \in \operatorname{int}(X + Y)$. We get a contradiction. In the second case,

$$B_\varepsilon(x + y) = B_\varepsilon(z) = B_\varepsilon(x) + y \subset \mathbb{R}^n \setminus [X + Y].$$

In this case as well, we obtain $z \notin \partial(X + Y)$ which contradicts our assumption $z \in \partial(X + Y)$. Thus, $z \in \partial X + \partial Y$. Theorem 3.8 is proved. □

It should be noted that the relation $\partial(X + Y) \supset \partial X + \partial Y$ may not hold. For example, consider the squares

$$X = \{(x, y) \in \mathbb{R}^2 \colon 1 \leq x \leq 4,\ 3 \leq y \leq 6\},$$

$$Y = \{(x, y) \in \mathbb{R}^2 \colon\ 4 \le x \le 5,\ 1 \le y \le 2\}.$$

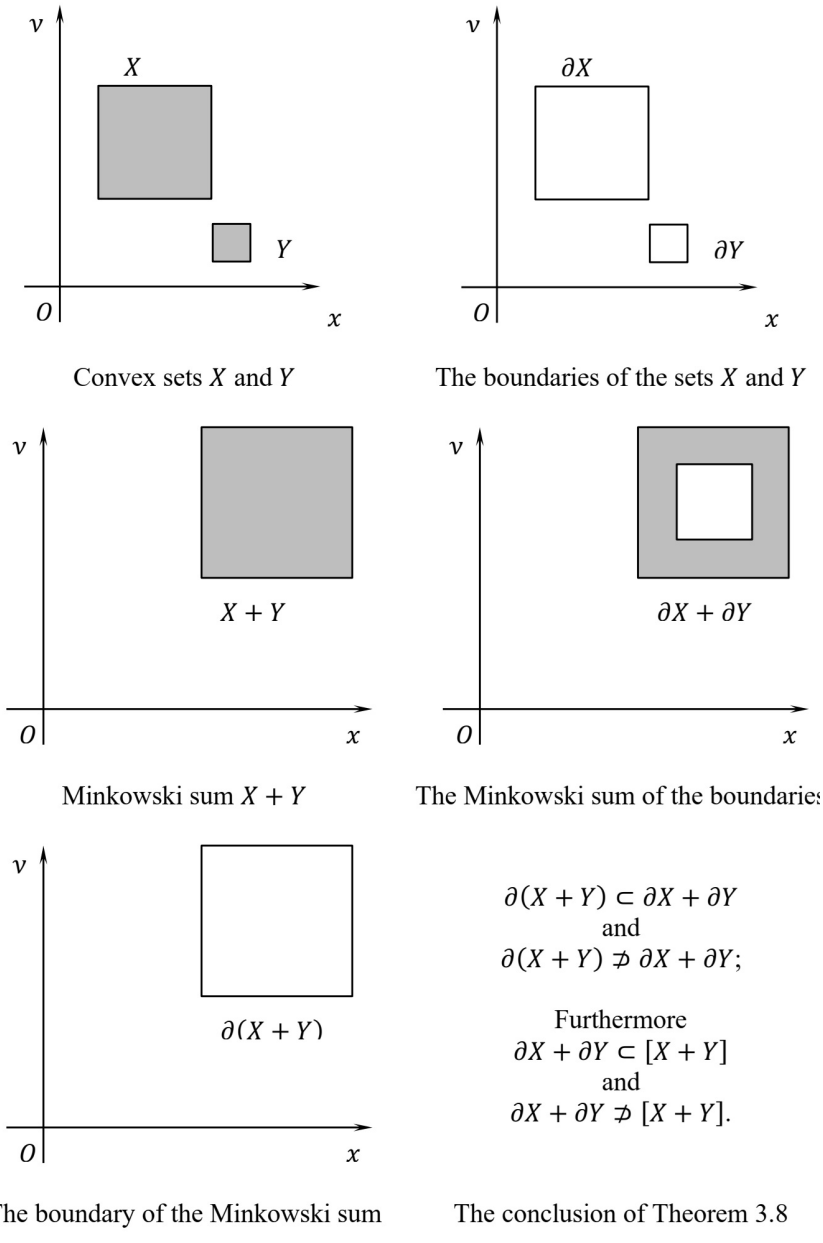


FIGURE 6.

It is clear that (see Fig. 6),

$$\partial X = \{(x, y) \in \mathbb{R}^2 \colon\ (x \in \{1, 4\},\ 3 \le y \le 6) \vee (1 \le x \le 4,\ y \in \{3, 6\})\},$$

$$\partial Y = \{(x, y) \in \mathbb{R}^2 \colon\ (x \in \{4, 5\},\ 1 \le y \le 2) \vee (4 \le x \le 5,\ y \in \{1, 2\})\},$$

$$\partial X + \partial Y = (X + Y) \setminus \{(x, y) \in \mathbb{R}^2 \colon\ 6 \le x \le 8,\ 5 \le y \le 7\},$$

$$X + Y = \{(x, y) \in \mathbb{R}^2\, 5 \le x \le 9,\ 4 \le y \le 8\},$$

$$\partial(X + Y) = \{(x, y) \in \mathbb{R}^2 \colon\ (x \in \{5, 9\},\ 4 \le y \le 8) \vee (5 \le x \le 9,\ y \in \{4\,8\})\}.$$

Thus,

$$\partial(X+Y) \subset \partial X + \partial Y$$

and

$$\partial(X+Y) \not\supset \partial X + \partial Y.$$

References

[1] Ivanov G. E., Weakly convex sets and their properties. Mathematical Notes, (2006) Vol. 79, Iss. 1, P. 55–78.

[2] Schneider R., Convex Bodies: The Brunn–Minkowski Theory. Cambridge University Press, Cambridge, (1993).

[3] Polovinkin E. S., Balashov M. V., Elements of Convex and Strongly Convex Analysis. Fizmatlit, Moscow, (2007).

[4] Mamatov M. S., Nuritdinov J. T., On some laws for calculating the Minkowski difference and sum. Bulletin of the Institute of Mathematics, (2020), Iss. 3, P. 49–59.

[5] Nuritdinov J. T., On the Minkowski difference of triangles. Bulletin of the Institute of Mathematics, (2021) Vol. 4, Iss. 6, P. 50–57.

[6] Mamatov M. S., Nuritdinov J. T., Turakulov K. S., Mamazhonov S. M., Geometric properties of the Minkowski operator. Bulletin of the Karaganda University. Mathematics Series, (2024) Vol. 11, Iss. 4, P. 127–137.

Zaitov Adilbek Atakhanovich,
Department of Mathematics and natural sciences,
Tashkent University of Architecture and Civil Engineering,
Tashkent, Uzbekistan,
e-mail: adilbek_zaitov@mail.ru

Nuritdinov Jalolxon,
Department of Mathematics, Kokand state university,
Kokand, Uzbekistan
e-mail: nuritdinovjt@gmail.com

Contents

Abdullaev A., Azamov A., Ruziboev M., Tilavov A. *On the accuracy of approximate solution scheme based on the Taylor formula for a Cauchy problem for ODEs* 5
Akrami S. E. *Schrödinger equation is a weak form of the classical Euler-Lagrange equation* 15
Amanova N. *On some existence theorems for quasilinear elliptic equations* . 24
Benaissa S., Messelmi F., Merouani A. *Solvability of a nonlinear elliptic problem involving p−Laplacian with Dirichlet condition* 33
Bouchikhi A., Mezroui S. *Eta quotients of level* 20 *and weight* 1 42
Chesneau C. *Bounds for a class of two-dimensional integrals involving mixed power-type singularities* 49
Fayazov K., Juraeva D. *Inverse problem for a system of mixed-type second-order partial differential equations* 59
Ganikhodjaev N. *Absorbing quadratic stochastic operators in genetics* 68
Hasanov A., Ergashev T., Tulakova Z. *On a mixed problem for a multi-dimensional elliptic equation with singular coefficients* 74
Ibragimov G., Muxammadjonov A., Qushaqov X. *An evasion game for an infinite system of ternary differential equations in the Hilbert space* l_2 83
Ibragimov G.I., Inomiddinov S.N., Ibrokhimjonov I.I. *Evasion differential game of one evader and multiple pursuers with changing energies* 90
Ibragimov M. *Geometric description of the unit ball of a neutral strongly facially symmetric space* 102
Kader S. *Integer Powers and Catalan- Mihăilescu Theorem* 111
Karimova N. *On the negativity of separable algorithmic representations of fields with additive inversions* 115
Khachiaa N. *Frames and operators on quaternionic Hilbert spaces* 121
Khayriev U., Nutfullaeva A., Hamroeva Z. *Optimization of the Euler-Maclaurin type quadrature formula in the Hilbert space of periodic functions* 136
Khusanbaev Ya.M. *Limit theorem for one scheme of particle evolution according to the critical Galton-Watson branching processes.* 145
Mahdi H. S., Rasouli S. H., Mohammadzadeh B. *On fixed point results in a new type of nonlinear contraction in* $\mathfrak{b}$*-metric spaces with applications to differential equations* 151
Muminov M., Jumaev Z. *Mixed piecewise constant argument and the Runge-Kutta method for numerically solving first-order differential equations* 159
Narmanov A., Parmonov H. *Geometry of orbits of Hamiltonian vector fields* 167
Normatov Z. *Anti-associative dendriform algebras* 174
Rakhmonov Z., Zaripova A. *Asymptotic Properties and Numerical Results of Solutions for a System of Degenerate Parabolic Equations with Nonlinear Boundary Conditions* 187
Usmanov Kh.I. *Application of the method of quadrature's with derivatives for solving problems of the Euler - Bernoulli beam* 199
Zaitov A., Nuritdinov J. *Geometric and topological properties of the Minkowski operator on convex sets* 208

Computer imposition: *A.F. Aliyev*

The journal was registered by the Press and Information Agency of the Republic of Uzbekistan on December 22, 2006. Register. No 0044.

Handed over to the set on 10/06/2026. Signed for printing on 15/06/2026
Format 60×84 1/16. Literary typeface. Offset printing.

V.I. Romanovskiy Institute of Mathematics, Uzbekistan Academy of Sciences 9 University st. 100174

Printed in a printing house "MERIT-PRINT"
Tashkent city, Yakkasaray district, Sh. Rustaveli street, 91