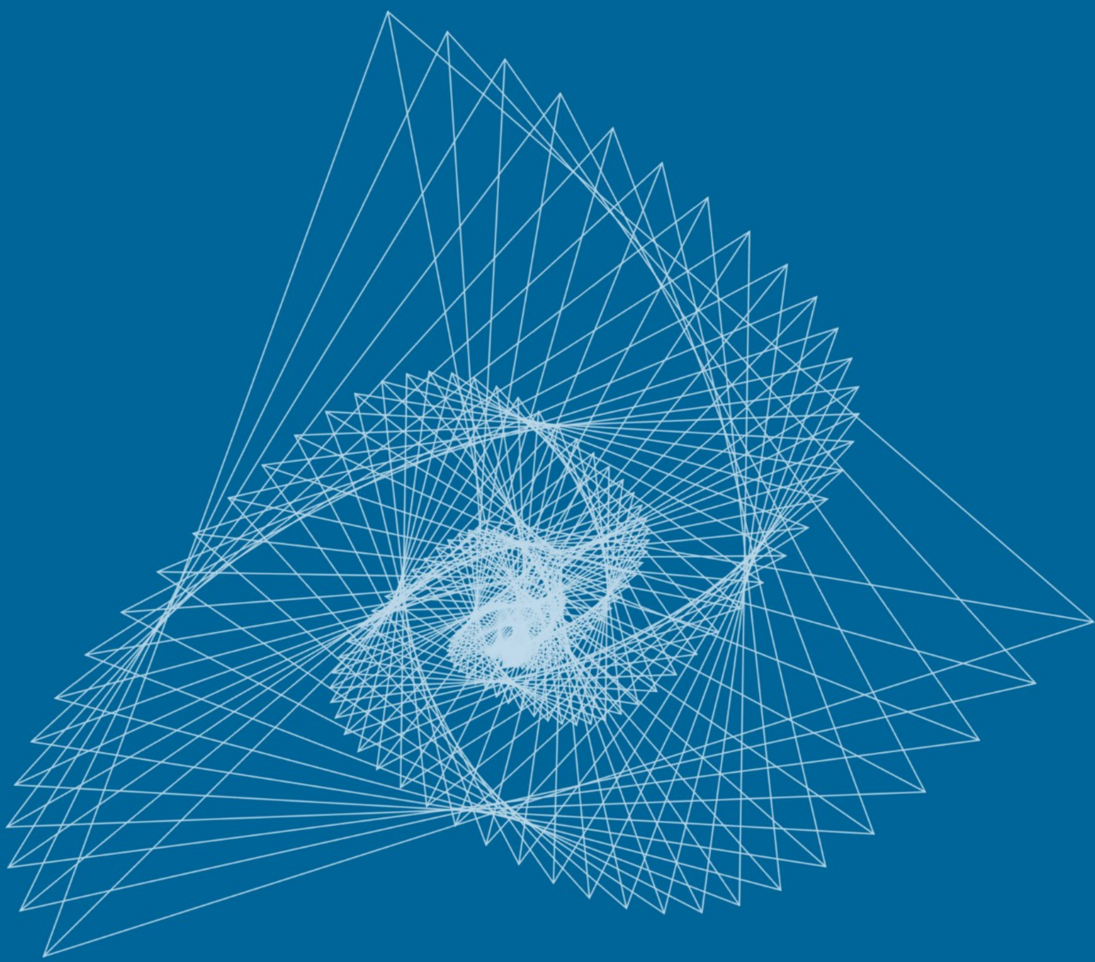
VOL. 7, NO. 1

URAL MATHEMATICAL JOURNAL

N.N. Krasovskii Institute of Mathematics and Mechanics of
the Ural Branch of Russian Academy of Sciences and
Ural Federal University named after the first President of Russia B.N.Yeltsin

ISSN: 2414-3952



UMJURAN.RU
UMJ

2021

ISSN 2414-3952
(Online)

Ural Mathematical Journal

Vol. 7, No. 1, 2021

https://umjuran.ru

Electronic Periodical Scientific Journal
Founded in 2015

The Journal is registered by the Federal Service for Supervision in the Sphere of Communication, Information Technologies and Mass Communications
Certificate of Registration of the Mass Media Эл № ФС77-61719 of 07.05.2015

Founders

N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N. Yeltsin

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990
Phone: +7 (343) 375-34-73 Fax: +7 (343) 374-25-81
Email: secretary@umjuran.ru
Web-site: https://umjuran.ru

TABLE OF CONTENTS

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 3–15

DOI: 10.15826/umj.2021.1.001

ON THE CHARACTERIZATION OF SCALING FUNCTIONS ON NON-ARCHEMEDEAN FIELDS

Ishtaq Ahmed[†], **Owias Ahmad**[††], **Neyaz Ahmad Sheikh**[†††]

National Institute of Technology,
Jammu and Kashmir, Srinagar-190006, India

[†]ishtiyaqahmadun@gmail.com, [††]siawoahmad@gmail.com, [†††]neyaznit@yahoo.co.in

Abstract: In real life application all signals are not obtained from uniform shifts; so there is a natural question regarding analysis and decompositions of these types of signals by a stable mathematical tool. This gap was filled by Gabardo and Nashed [11] by establishing a constructive algorithm based on the theory of spectral pairs for constructing non-uniform wavelet basis in $L^2(\mathbb{R})$. In this setting, the associated translation set $\Lambda = \{0, r/N\} + 2\,\mathbb{Z}$ is no longer a discrete subgroup of $\mathbb{R}$ but a spectrum associated with a certain one-dimensional spectral pair and the associated dilation is an even positive integer related to the given spectral pair. In this paper, we characterize the scaling function for non-uniform multiresolution analysis on local fields of positive characteristic (LFPC). Some properties of wavelet scaling function associated with non-uniform multiresolution analysis (NUMRA) on LFPC are also established.

Keywords: Scaling function, Fourier transform, Local field, NUMRA

1. Introduction

Multiresolution analysis (MRA) is an important mathematical tool since it provides a natural framework for understanding and constructing discrete wavelet systems. The concept of MRA provides a natural framework for understanding and constructing discrete wavelet systems. Multiresolution analysis is an increasing family of closed spaces $\{V_j : j \in \mathbb{Z}\}$ of $L^2(\mathbb{R})$ such that $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$ and $\bigcup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$ which satisfies $f \in V_j$ if and only if $f(2\cdot) \in V_{j+1}$. Moreover, there exists a function $\varphi \in V_0$ such that the collection of integer translates of the function φ, $\{\varphi(\cdot - k) : k \in \mathbb{Z}\}$, represents a complete orthonormal system for V_0. The function φ is called scaling function or father wavelet. The concept of multiresolution analysis has been extended in various ways in recent years. These concepts are generalized to $L^2(\mathbb{R}^d)$, to lattices different from $\mathbb{Z}^d$, allowing the subspaces of MRA to be generated by Riesz basis instead of orthonormal basis, admitting a finite number of scaling functions, replacing the dilation factor 2 by an integer $M \geq 2$ or by an expansive matrix $A \in GL_d(\mathbb{R})$ as long as $A \subset A\mathbb{Z}^d$. All these concepts are developed on regular lattices, that is the translation set is always a group. Recently, Gabardo and Nashed [11] considered a generalization of Mallat's [21] celebrated theory of MRA based on spectral pairs, in which the translation set acting on the scaling function associated with the MRA to generate the subspace V_0 is no longer a group, but is the union of $\mathbb{Z}$ and a translate of $\mathbb{Z}$. Based on one-dimensional spectral pairs, Gabardo and Yu [12] considered sets of nonuniform wavelets in $L^2(\mathbb{R})$. In the heart of any MRA, there lies the concept of scaling functions. Cifuentes et al. [10] characterized the scaling function of MRA in a general settings. The multiresolution analysis whose scaling functions are characteristic functions some elementary properties of MRA of $L^2(\mathbb{R}^n)$ are established by Madych [20]. Zhang [26] studied scaling functions of standard MRA and wavelets. Zhang [26] characterized support of the Fourier transform of scaling functions.

The theory of wavelets, wavelet frames, multiresolution analysis, Gabor frames on local fields of positive characteristics (LFPC) are extensively studied by many researchers including Benedetto, Behera and Jahan, Ahmed and Neyaz, Ahmad and Shah, Jiang, Li and Ji in the references [1–4, 7–9, 13, 19, 22, 24] but still more concepts required to be studied for its enhancement on LFPC. Albeverio, Kozyrev, Khrennikov, Shelkovich, Skopina and their collaborators also established the theory of MRA and wavelets on the p-adic field $\mathbb{Q}_p$ in a series of papers [5, 6, 14–18], where $\mathbb{Q}_p$ is a local field of characteristic 0. Recently, Shah and Abdullah [23] have generalized the concept of multiresolution analysis on Euclidean spaces $\mathbb{R}^n$ to nonuniform multiresolution analysis on local fields of positive characteristic, in which the translation set acting on the scaling function associated with the multiresolution analysis to generate the subspace V_0 is no longer a group, but is the union of $\mathcal{Z}$ and a translate of $\mathcal{Z}$, where $\mathcal{Z} = \{u(n) : n \in \mathbb{N}_0\}$ is a complete list of (distinct) coset representation of the unit disc $\mathfrak{D}$ in the locally compact Abelian group $\mathbb{K}^+$. More precisely, this set is of the form $\Lambda = \{0, r/N\} + \mathcal{Z}$, where $N \geq 1$ is an integer and r is an odd integer such that r and N are relatively prime. They call this a *nonuniform multiresolution analysis* on local fields of positive characteristic. Inspired by the work of Shah and Abdullah [23], we in this paper establish the characterization of scaling function for nonuniform multiresolution on local fields of positive characteristic. Some properties of wavelet scaling functions associated with NUMRA on LFPC are established.

The remainder of the paper is structured as follows. In Section 2, we discuss preliminary results on local fields as well as some definitions and auxiliary results. Section 3 is devoted to the characterization of scaling function associated with nonuniform multiresolution analysis on LFPC.

2. Preliminaries on local fields

2.1. Local fields

A local field $\mathbb{K}$ is a locally compact, non-discrete and totally disconnected field. If it is of characteristic zero, then it is a field of p-adic numbers $\mathbb{Q}_p$ or its finite extension. If $\mathbb{K}$ is of positive characteristic, then $\mathbb{K}$ is a field of formal Laurent series over a finite field $GF(p^c)$. If $c = 1$, it is a p-series field, while for $c \neq 1$, it is an algebraic extension of degree c of a p-series field. Let $\mathbb{K}$ be a fixed local field with the ring of integers

$$\mathfrak{D} = \{x \in K : |x| \leq 1\}.$$

Since K^+ is a locally compact Abelian group, we choose a Haar measure dx for K^+. The field K is locally compact, non-trivial, totally disconnected and complete topological field endowed with non–Archimedean norm $|\cdot| : \mathbb{K} \to \mathbb{R}^+$ satisfying

(a) $|x| = 0$ if and only if $x = 0$;

(b) $|x\,y| = |x||y|$ for all $x, y \in \mathbb{K}$;

(c) $|x + y| \leq \max\{|x|, |y|\}$ for all $x, y \in \mathbb{K}$.

Property (c) is called the ultrametric inequality. Let $\mathfrak{B} = \{x \in \mathbb{K} : |x| < 1\}$ be the prime ideal of the ring of integers $\mathfrak{D}$ in $\mathbb{K}$. Then, the residue space $\mathfrak{D}/\mathfrak{B}$ is isomorphic to a finite field $GF(q)$, where $q = p^c$ for some prime p and $c \in \mathbb{N}$. Since K is totally disconnected and $\mathfrak{B}$ is both prime and principal ideal, so there exist a prime element $\mathfrak{p}$ of $\mathbb{K}$ such that $\mathfrak{B} = \langle\mathfrak{p}\rangle = \mathfrak{p}\mathfrak{D}$.

Let

$$\mathfrak{D}^* = \mathfrak{D} \setminus \mathfrak{B} = \{x \in \mathbb{K} : |x| = 1\}.$$

Clearly, $\mathfrak{D}^*$ is a group of units in $\mathbb{K}^*$ and if $x \neq 0$, then can write $x = \mathfrak{p}^n y, y \in \mathfrak{D}^*$. Moreover, if $\mathcal{U} = \{a_m : m = 0, 1, \ldots, q-1\}$ denotes the fixed full set of coset representatives of $\mathfrak{B}$ in $\mathfrak{D}$, then every element $x \in K$ can be expressed uniquely as

$$x = \sum_{\ell=k}^{\infty} c_\ell \, \mathfrak{p}^\ell, \quad \text{with} \quad c_\ell \in \mathcal{U}.$$

Recall that $\mathfrak{B}$ is compact and open, so each fractional ideal

$$\mathfrak{B}^k = \mathfrak{p}^k \mathfrak{D} = \left\{ x \in K : |x| < q^{-k} \right\}$$

is also compact and open and is a subgroup of K^+. We use the notation in Taibleson's book [25]. In the rest of this paper, we use the symbols $\mathbb{N}, \mathbb{N}_0$ and $\mathbb{Z}$ to denote the sets of natural, non-negative integers and integers, respectively.

Let χ be a fixed character on K^+ that is trivial on $\mathfrak{D}$ but non-trivial on $\mathfrak{B}^{-1}$. Therefore, χ is constant on cosets of $\mathfrak{D}$ so if $y \in \mathfrak{B}^k$, then $\chi_y(x) = \chi(y, x), x \in K$. Suppose that χ_u is any character on K^+, then the restriction $\chi_u | \mathfrak{D}$ is a character on $\mathfrak{D}$. Moreover, as characters on $\mathfrak{D}, \chi_u = \chi_v$ if and only if $u - v \in \mathfrak{D}$. Hence, if $\{u(n) : n \in \mathbb{N}_0\}$ is a complete list of distinct coset representative of $\mathfrak{D}$ in K^+, then, as it was proved in [25], the set $\left\{\chi_{u(n)} : n \in \mathbb{N}_0\right\}$ of distinct characters on $\mathfrak{D}$ is a complete orthonormal system on $\mathfrak{D}$.

We now impose a natural order on the sequence $\{u(n)\}_{n=0}^{\infty}$. We have $\mathfrak{D}/\mathfrak{B} \cong GF(q)$ where $GF(q)$ is a c-dimensional vector space over the field $GF(p)$. We choose a set

$$\{1 = \zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_{c-1}\} \subset \mathfrak{D}^*$$

such that span $\{\zeta_j\}_{j=0}^{c-1} \cong GF(q)$. For $n \in \mathbb{N}_0$ satisfying

$$0 \le n < q, \quad n = a_0 + a_1 p + \cdots + a_{c-1} p^{c-1}, \quad 0 \le a_k < p, \quad k = 0, 1, \ldots, c-1,$$

we define

$$u(n) = (a_0 + a_1 \zeta_1 + \cdots + a_{c-1} \zeta_{c-1}) \, \mathfrak{p}^{-1}.$$

Also, for

$$n = b_0 + b_1 q + b_2 q^2 + \cdots + b_s q^s, \quad n \in \mathbb{N}_0, \quad 0 \le b_k < q, \quad k = 0, 1, 2, \ldots, s,$$

we set

$$u(n) = u(b_0) + u(b_1) \mathfrak{p}^{-1} + \cdots + u(b_s) \mathfrak{p}^{-s}.$$

This defines $u(n)$ for all $n \in \mathbb{N}_0$. In general, it is not true that $u(m+n) = u(m) + u(n)$. But, if $r, k \in \mathbb{N}_0$ and $0 \le s < q^k$, then

$$u(rq^k + s) = u(r) \mathfrak{p}^{-k} + u(s).$$

Further, it is also easy to verify that $u(n) = 0$ if and only if $n = 0$ and

$$\{u(\ell) + u(k) : k \in \mathbb{N}_0\} = \{u(k) : k \in \mathbb{N}_0\}$$

for a fixed $\ell \in \mathbb{N}_0$. Hereafter we use the notation $\chi_n = \chi_{u(n)}$, $n \ge 0$.

Let the local field $\mathbb{K}$ be of characteristic $p > 0$ and $\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_{c-1}$ be as above. We define a character χ on K as follows:

$$\chi(\zeta_\mu \mathfrak{p}^{-j}) = \begin{cases} \exp(2\pi i/p), & \mu = 0 \quad \text{and} \quad j = 1, \\ 1, & \mu = 1, \ldots, c-1 \quad \text{or} \quad j \ne 1. \end{cases}$$

2.2. Fourier transforms on local fields

The Fourier transform of $f \in L^1(K)$ is denoted by $\hat{f}(\xi)$ and defined by

$$\mathcal{F}\{f(x)\} = \hat{f}(\xi) = \int_K f(x)\overline{\chi_\xi(x)}\, dx.$$

It is noted that

$$\hat{f}(\xi) = \int_K f(x)\, \overline{\chi_\xi(x)} dx = \int_K f(x)\chi(-\xi x)\, dx.$$

The properties of Fourier transforms on local field $\mathbb{K}$ are much similar to those of on the classical field $\mathbb{R}$. In fact, the Fourier transform on local fields of positive characteristic have the following properties:

- The map $f \to \hat{f}$ is a bounded linear transformation of $L^1(\mathbb{K})$ into $L^\infty(\mathbb{K})$, and $\|\hat{f}\|_\infty \leq \|f\|_1$.
- If $f \in L^1(\mathbb{K})$, then $\hat{f}$ is uniformly continuous.
- If $f \in L^1(\mathbb{K}) \cap L^2(\mathbb{K})$, then $\|\hat{f}\|_2 = \|f\|_2$.

The Fourier transform of a function $f \in L^2(\mathbb{K})$ is defined by

$$\hat{f}(\xi) = \lim_{k\to\infty} \hat{f}_k(\xi) = \lim_{k\to\infty} \int_{|x|\leq q^k} f(x)\overline{\chi_\xi(x)}\, dx,$$

where $f_k = f\,\Phi_{-k}$ and Φ_k is the characteristic function of $\mathfrak{B}^k$. Furthermore, if $f \in L^2(\mathfrak{D})$, then we define the Fourier coefficients of f as

$$\hat{f}\big(u(n)\big) = \int_{\mathfrak{D}} f(x)\overline{\chi_{u(n)}(x)}\, dx.$$

The series

$$\sum_{n\in\mathbb{N}_0} \hat{f}\big(u(n)\big)\chi_{u(n)}(x)$$

is called the Fourier series of f. From the standard L^2-theory for compact Abelian groups, we conclude that the Fourier series of f converges to f in $L^2(\mathfrak{D})$ and Parseval's identity holds:

$$\|f\|_2^2 = \int_{\mathfrak{D}} \big|f(x)\big|^2 dx = \sum_{n\in\mathbb{N}_0} \Big|\hat{f}\big(u(n)\big)\Big|^2.$$

3. Nonuniform MRA on local fields

Definition 1. *For an integer $N \geq 1$ and an odd integer r with $1 \leq r \leq qN-1$ such that r and N are relatively prime, we define*

$$\Lambda = \left\{0, \frac{u(r)}{N}\right\} + \mathcal{Z}$$

and

$$\Delta_N = \left\{u(m)N + \mathfrak{p}u(j):\ m \in \mathbb{Z},\ 0 \leq j \leq N-1\right\},$$

where

$$\mathcal{Z} = \left\{u(n) : n \in \mathbb{N}_0\right\}.$$

It is easy to verify that Λ is not a group on local field $\mathbb{K}$, but is the union of $\mathcal{Z}$ and a translate of $\mathcal{Z}$.

Following is the definition of nonuniform multiresolution analysis (NUMRA) on local fields of positive characteristic given by Shah and Abdullah [23].

Definition 2. *For an integer $N \geq 1$ and an odd integer r with $1 \leq r \leq qN-1$ such that r and N are relatively prime, an associated NUMRA on local field $\mathbb{K}$ of positive characteristic is a sequence of closed subspaces $\{V_j : j \in \mathbb{Z}\}$ of $L^2(\mathbb{K})$ such that the following properties hold:*

(a) $V_j \subset V_{j+1}$ *for all* $j \in \mathbb{Z}$;

(b) $\bigcup_{j\in\mathbb{Z}} V_j$ *is dense in* $L^2(\mathbb{K})$;

(c) $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$;

(d) $f(\cdot) \in V_j$ *if and only if* $f(\mathfrak{p}^{-1}N\cdot) \in V_{j+1}$ *for all* $j \in \mathbb{Z}$;

(e) *There exists a function φ in V_0 such that $\{\varphi(\cdot - \lambda) : \lambda \in \Lambda\}$, is a complete orthonormal basis for V_0.*

It is worth noticing that, when $N = 1$, one recovers the definition of an MRA on local fields of positive characteristic $p > 0$. When, $N > 1$, the dilation is induced by $\mathfrak{p}^{-1}N$ and $|\mathfrak{p}^{-1}| = q$ ensures that $qN\Lambda \subset \mathcal{Z} \subset \Lambda$. For every $j \in \mathbb{Z}$, define W_j to be the orthogonal complement of V_j in V_{j+1}.

Then we have

$$V_{j+1} = V_j \oplus W_j \quad \text{and} \quad W_\ell \perp W_{\ell'} \quad \text{if} \quad \ell \neq \ell'.$$

It follows that for $j > J$,

$$V_j = V_J \oplus \bigoplus_{\ell=0}^{j-J-1} W_{j-\ell}\,,$$

where all these subspaces are orthogonal. By virtue of condition (b) in the Definition 2, this implies

$$L^2(\mathbb{K}) = \bigoplus_{j\in\mathbb{Z}} W_j,$$

a decomposition of $L^2(\mathbb{K})$ into mutually orthogonal subspaces.

As in the standard scheme, one expects the existence of $qN-1$ number of functions so that their translation by elements of Λ and dilations by the integral powers of $\mathfrak{p}^{-1}N$ form an orthonormal basis for $L^2(\mathbb{K})$.

Let a and b be any two fixed elements in $\mathbb{K}$. Then, for any prime $\mathfrak{p}$ and $m, n \in \mathbb{N}_0$, let $D_\mathfrak{p}, T_{u(n)a}$ and $E_{u(m)b}$ be the unitary operators acting on $f \in L^2(\mathbb{K})$ defined by:

$$\begin{aligned} T_{u(n)a} f(x) &= f\big(x - u(n)a\big), \qquad \text{(Translation)},\\ E_{u(m)b} f(x) &= \chi\big(u(m)bx\big) f(x), \qquad \text{(Modulation)},\\ D_\mathfrak{p} f(x) &= \sqrt{qN} f\left(\mathfrak{p}^{-1}Nx\right), \qquad \text{(Dilation)}. \end{aligned}$$

Then for any $f \in L^2(K)$, the following results can easily be verified:

$$\begin{aligned} \mathcal{F}\big\{T_{u(n)a} f(x)\big\} &= E_{-u(n)a} \mathcal{F}\big\{f(x)\big\},\\ \mathcal{F}\big\{E_{u(m)b} f(x)\big\} &= T_{u(m)b} \mathcal{F}\big\{f(x)\big\},\\ \mathcal{F}\big\{D_{\mathfrak{p}^j} f(x)\big\} &= D_{\mathfrak{p}^{-j}} \mathcal{F}\big\{f(x)\big\},\\ D_{\mathfrak{p}^j} T_{u(n)a} &= T_{(qN)^{-j}u(n)a} D_{\mathfrak{p}^j}. \end{aligned}$$

We state the following lemmas which will be very useful in establishing the results and whose proof can be found in [23].

Lemma 1. *For an integer $N \geq 1$ and an odd integer r with $1 \leq r \leq qN-1$ such that r and N are relatively prime. Let $\varphi \in L^2(\mathbb{K})$ with $\|\varphi\|^2 = 1$, then*

(i) *the family $\{\varphi(\xi-\lambda) : \lambda \in \Lambda\}$ is an orthonormal system for fixed r if and only if*

$$\sum_{k\in\mathbb{N}_0} |\widehat{\varphi}\left(\xi + \mathfrak{p}u(k)\right)|^2 = q \quad a.e. \quad \xi \in \mathbb{K}$$

and

$$\sum_{k\in\mathbb{N}_0} \overline{\chi\left(\frac{u(r)}{N}u(k)\right)} |\widehat{\varphi}\left(\xi + \mathfrak{p}u(k)\right)|^2 = 0 \quad a.e. \quad \xi \in \mathbb{K};$$

(ii) *the family $\{\varphi(\xi-\lambda) : \lambda \in \Delta_N\}$ is an orthonormal system for every odd integer r if and only if*

$$|\widehat{\varphi}(\xi-\gamma)|^2 = 1, \quad a.e. \quad \xi \in \mathbb{K}.$$

Lemma 2. *Let (V_j, φ) be non-uniform multiresolution analysis, where*

$$V_0 = \overline{\operatorname{span}}\,\{\varphi(x-\lambda) : \lambda \in \Lambda\}.$$

Then the necessary and sufficient condition for the existence of associated wavelets is

$$\sum_{\gamma\in\Delta_N} |\widehat{\varphi}(\xi-\gamma)|^2 = 1 \quad a.e. \quad \xi \in \mathbb{K}.$$

Lemma 3. *Let $\mathcal{S} \subset \mathbb{K}$ be measurable and $\Lambda_0 = \{0, u(a)\} + \mathcal{Z}$. Then $(\mathcal{S}, \Lambda_0)$ is a spectral pair if and only if there exist an integer $N \geq 1$ and an odd integer r with $1 \leq r \leq qN-1$, such that N and r are relatively prime , $a = r/N$ and*

$$\sum_{j=0}^{N-1} \delta_{j/2} \star \sum_{n\in\mathbb{N}_0} \delta_{nN} \star \mathbf{\Phi}_{\mathcal{S}} = 1.$$

4. Characterization of scaling functions on LFPC

In this section, we establish the characterization of scaling functions associated with nonuniform multiresolution analysis on LFPC. We also provide the sufficient condition for the frequency band of the scaling function on LFPC.

Theorem 1. *A nonzero function $\varphi \in L^2(\mathbb{K})$ is a scaling function for wavelet NUMRA if and only if the following conditions are satisfied*

(i) $\sum_{\gamma\in\Delta_N} |\widehat{\varphi}(\xi-\gamma)|^2 = 1 \quad a.e. \quad \xi \in \mathbb{K};$

(ii) $\lim_{j\to\infty} \left|\widehat{\varphi}(\mathfrak{p}^{-1}N)^j\xi\right|^2 = 1 \quad a.e. \quad \xi \in q^2\mathfrak{D};$

(iii) *there exist functions $m^1(\xi), m^2(\xi)$ locally integrable, q-periodic functions such that*

$$\widehat{\varphi}(\mathfrak{p}^{-1}N\xi) = m(\xi)\widehat{\varphi}(\xi) \quad a.e. \quad \xi \in \mathbb{K},$$

where

$$m(\xi) = m^1(\xi) + \overline{\chi\left(\frac{u(r)}{N}\xi\right)} m^2(\xi).$$

P r o o f. Suppose $\psi \in L^2(\mathbb{K})$ is a scaling function for wavelet NUMRA, say $\{V_j, \varphi\}_{j\in\mathbb{Z}}$. Then by Lemma 2, we must have

$$\sum_{\gamma\in\Delta_N} |\widehat{\varphi}(\xi-\gamma)|^2 = 1 \quad \text{a.e.} \quad \xi \in \mathbb{K}. \tag{4.1}$$

This gives (i). Since $\varphi \in V_0$, we have $D_{\mathfrak{p}^{-1}}\varphi \in V_{-1} \subseteq V_0$. Thus we can write

$$D_{\mathfrak{p}^{-1}}\varphi = \sum_{\lambda\in\Lambda} a_\lambda T_\lambda \varphi.$$

Taking the Fourier transform of both sides, we get

$$D_{\mathfrak{p}}\widehat{\varphi} = \sum_{\lambda\in\Lambda} a_\lambda E_{-\lambda}\widehat{\varphi}.$$

So we can write,

$$\widehat{\varphi}(\mathfrak{p}^{-1}N\gamma) = m(\gamma)\widehat{\varphi}(\gamma),$$

where

$$m(\gamma) = m^1(\gamma) + \overline{\chi(\frac{u(r)}{N}\gamma)}m^2(\gamma)$$

and m^1, m^2 are q- periodic and locally integrable functions. This proves (iii).

Next we show that (ii) holds. Let $f \in L^2(\mathbb{K})$ be such that $\widehat{f}(\gamma) = \Phi_{q^2\mathfrak{D}}(\gamma)$. Then

$$\|f\|^2 = \|\widehat{f}\|^2 = q.$$

As (V_j, φ) is NUMRA so if P_j is orthogonal projection onto V_j, we must have

$$\|f - P_j f\|^2 \to 0 \quad \text{as} \quad j \to \infty.$$

That is

$$\|P_j f\| \to \|f\| \quad \text{as} \quad j \to \infty.$$

Since $\{T_\lambda\varphi\}_{\lambda\in\Lambda}$ is an orthonormal bases for V_0 so $\{D_{\mathfrak{p}^j}T_\lambda\varphi\}_{\lambda\in\Lambda}$ is an orthonormal basis for V_j. Thus

$$\|P_j f\|^2 = \sum_{\lambda\in\Lambda} |\langle f, D_{\mathfrak{p}^j}T_\lambda\varphi\rangle|^2 \to \frac{1}{q} \quad \text{a.e.} \quad j \to \infty \tag{4.2}$$

$$\begin{aligned}
\sum_{\lambda\in\Lambda} |\langle f, D_{\mathfrak{p}^j}T_\lambda\varphi\rangle|^2 &= \sum_{\lambda\in\mathcal{Z}} |\langle f, D_{\mathfrak{p}^j}T_\lambda\varphi\rangle|^2 + \sum_{\lambda\in(u(r)/N+\mathcal{Z})} |\langle f, D_{\mathfrak{p}^j}T_\lambda\varphi\rangle|^2 \\
&= \sum_{\lambda\in\mathcal{Z}} |\widehat{f}, \widehat{D_{\mathfrak{p}^j}T_\lambda\varphi}|^2 + \sum_{\lambda\in(u(r)/N+\mathcal{Z})} |\widehat{f}, \widehat{D_{\mathfrak{p}^j}T_\lambda\varphi}|^2 \\
&= \sum_{k\in\mathbb{N}_0} \left| \int_{\mathbb{K}} (qN)^{-j/2}\widehat{f}(\gamma)\chi_{u(k)}\Big(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\Big)\overline{\widehat{\varphi}\Big(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\Big)}d\gamma\right|^2 \\
&+ \sum_{k\in\mathbb{N}_0} \left| \int_{\mathbb{K}} (qN)^{-j/2}\widehat{f}(\gamma)\chi_{u(k)}\Big(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\Big(\frac{u(r)}{N} + \mathfrak{p}u(k)\Big)\Big)\overline{\widehat{\varphi}\Big(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\Big)}d\gamma\right|^2
\end{aligned}$$

$$= \sum_{k\in\mathbb{N}_0} \left| \int_{q^2\mathfrak{D}} (qN)^{-j/2} \chi_{u(k)} \left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\right) \overline{\widehat{\varphi}\left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\right)} d\gamma \right|^2$$

$$= \sum_{k\in\mathbb{N}_0} \left| \int_{q^2\mathfrak{D}} (qN)^{-j/2} \chi_{u(k)} \left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\left(\frac{u(r)}{N} + \mathfrak{p}u(k)\right)\right) \overline{\widehat{\varphi}\left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\right)} d\gamma \right|^2$$

Putting

$$\frac{\gamma}{(qN)^j} = \eta,$$

we obtain

$$\begin{aligned}
\sum_{\lambda\in\Lambda} |\langle f, D_{\mathfrak{p}^j} T_\lambda \varphi\rangle|^2 &= \frac{(qN)^j}{2}\Bigg\{ \sum_{k\in\mathbb{N}_0} \left| \int_{(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D}} \sqrt{q}\chi_{u(k)}(\mathfrak{p}^{-1}\eta)\overline{\widehat{\varphi}(\eta)} d\eta \right|^2 \\
&\quad + \sum_{k\in\mathbb{N}_0} \left| \int_{(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D}} \sqrt{q}\chi_{u(k)}(\mathfrak{p}^{-1}\eta)\overline{\widehat{\varphi}(\eta)} d\eta \right|^2 \Bigg\} \\
&= \frac{(qN)^j}{2}\Bigg\{ \sum_{k\in\mathbb{N}_0} \left| \int_{q^2\mathfrak{D}} \mathbf{\Phi}_{(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D}} \sqrt{q}\chi_{u(k)}(\mathfrak{p}^{-1}\eta)\overline{\widehat{\varphi}(\eta)} d\eta \right|^2 \\
&\quad + \sum_{k\in\mathbb{N}_0} \left| \int_{q^2\mathfrak{D}} \mathbf{\Phi}_{(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D}} \chi\big(\frac{u(r)}{N}\eta\big) \sqrt{q}\chi_{u(k)}(\mathfrak{p}^{-1}\eta)\overline{\widehat{\varphi}(\eta)} d\eta \right|^2 \Bigg\},
\end{aligned}$$

because $(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D} \subseteq q^2\mathfrak{D}$, for any $j \geq 0$. Therefore from (4.2) and from the fact that $\{\sqrt{q}\chi_{u(k)}(\mathfrak{p}^{-1}\eta)\}$ is an orthonormal basis for $L^2(q\mathfrak{D})$, we get

$$\sum_{\lambda\in\Lambda} |\langle f, D_{\mathfrak{p}^j} T_\lambda \varphi\rangle|^2 = (qN)^j \int_{(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D}} |\widehat{\varphi}(\eta)|^2 d\eta \to \frac{1}{q}\ j \to \infty.$$

Putting $\mu = (qN)^j\eta$, we get

$$\int_{q^2\mathfrak{D}} |\widehat{\varphi}(\mathfrak{p}^{-1}N)^j\mu|^2 d\mu \to \frac{1}{q} \quad \text{as} \quad j \to \infty. \tag{4.3}$$

Let

$$h(\xi) = \lim_{j\to\infty} |\widehat{\varphi}(\mathfrak{p}^{-1}N)^j\xi|^2.$$

Then

$$0 \leq h(\xi) \leq 1 \quad \text{a.e.} \quad \xi \in q^2\mathfrak{D}.$$

Indeed for any fixed $j \in \mathbb{Z}$ by using (4.1), we have

$$0 \leq |\widehat{\varphi}(\mathfrak{p}^{-1}N)^j\xi|^2 \leq 1 \quad \text{a.e.} \quad \xi \in q^2\mathfrak{D}.$$

This gives

$$0 \leq h(\xi) = \lim_{j\to\infty} |\widehat{\varphi}(\mathfrak{p}^{-1}N)^j\xi|^2 \leq 1 \quad \text{a.e.} \quad \xi \in q^2\mathfrak{D}.$$

Now invoking the Lesbesgue-dominated convergence theorem, we obtain

$$\lim_{j\to\infty} \int_{q^2\mathfrak{D}} |\widehat{\varphi}((\mathfrak{p}^{-1}N)^j)\mu|^2 d\mu = \int_{q^2\mathfrak{D}} \lim_{j\to\infty} |\widehat{\varphi}((\mathfrak{p}^{-1}N)^j)\mu|^2 d\mu = \frac{1}{q}.$$

Thus

$$\int_{q^2\mathfrak{D}} h(\xi) d\xi = \frac{1}{q} = \int_{q^2\mathfrak{D}} 1 d\xi.$$

That is

$$\int_{q^2\mathfrak{D}}(1-h(\xi))d\xi = 0,$$

so by using

$$0 \le h(\xi) \le 1 \quad \text{a.e.} \quad \xi \in q^2\mathfrak{D},$$

we get $h(\xi)=1$ a.e. $\xi \in q^2\mathfrak{D}$. Hence (ii) is proved.

Conversely, let $\varphi \in L^2(\mathbb{K})$ satisfying (i)–(iii). We define closed subspaces V_j of $L^2(\mathbb{K})$ in the following way.

For $j=0$ let $V_j = \overline{\text{span}}\{\varphi(\xi-\lambda) : \lambda \in \Lambda\}$ and for $j \neq 0$ let $V_j = \left\{f : f((\mathfrak{p}^{-1}N)^{-j}\xi) \in V_0\right\}$. We will show (V_j, φ) forms wavelet NUMRA. Using Lemma 1, the sequence $\{T_\lambda \varphi\}_{\lambda\in\Lambda}$ is an orthonormal basis for V_0.

By definition of V_j, it can be easily shown that $f(\gamma) \in V_j$ if and only if

$$f((\mathfrak{p}^{-1}N)\gamma) \in V_{j+1},$$

which clearly implies $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$. To prove $V_j \subseteq V_{j+1}$, it is sufficient to show that $V_0 \subseteq V_1$. First we show that

$$V_j = \left\{ f \in L^2(\mathbb{K}) : \widehat{f}((\mathfrak{p}^{-1}N)^j\gamma) = (m_j^1(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} m_j^2(\gamma))\widehat{\varphi}(\gamma) \right\}, \tag{4.4}$$

where m_j^1, m_j^2 are locally integrable, q-periodic functions. Let $f \in V_j$, then

$$\frac{1}{(qN)^{j/2}} D_{\mathfrak{p}^{-j}} f(\gamma) \in V_0,$$

as $\{T_\lambda\varphi\}_{\lambda\in\Lambda}$ is an orthonormal basis for V_0, so there exist $\{c_\lambda^j\} \in \ell^2(\mathbb{N}_0)$ such that

$$\frac{1}{(qN)^{j/2}} D_{\mathfrak{p}^{-j}} f(\gamma) = \sum_{\lambda\in\Lambda} c_\lambda^j T_\lambda \varphi.$$

On taking Fourier transform of both sides, we obtain

$$\widehat{f}((\mathfrak{p}^{-1}N)^j\gamma) = \sum_{\lambda\in\Lambda} c_\lambda^j \overline{\chi_\lambda(\mathfrak{p}^{-1}\gamma)}\widehat{\varphi}(\gamma) = \left\{ m_j^1(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} m_j^2(\gamma)\right\}\widehat{\varphi}(\gamma),$$

where m_j^1 and m_j^2 are locally integrable and q-periodic functions. If $f \in L^2(\mathbb{K})$ satisfies

$$\widehat{f}((\mathfrak{p}^{-1}N)^j\gamma) = \left\{ m_j^1(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} m_j^2(\gamma)\right\}\widehat{\varphi}(\gamma)$$

for some m_j^1 and m_j^2 are locally integrable and q-periodic functions, then we can write

$$\widehat{f}((\mathfrak{p}^{-1}N)^j\xi) = \left\{ \sum_{k\in\mathbb{Z}} c_k^j \overline{\chi_{u(k)}(\mathfrak{p}^{-1}\gamma)} + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} \sum_{k\in\mathbb{N}_0} d_k^j \overline{\chi_{u(k)}(\mathfrak{p}^{-1}\gamma)} \right\}\widehat{\varphi}(\gamma)$$

for some scalars $\{c_k^j\}$ and $\{d_k^j\}_{k\in\mathbb{N}_0} \in \ell^2(\mathbb{N}_0)$. Therefore

$$\frac{D_{\mathfrak{p}^j}\widehat{f}(\gamma)}{(qN)^{j/2}} = \sum_{\lambda\in\Lambda} l_\lambda^j \overline{\chi_{u(k)}(\mathfrak{p}^{-1}\gamma)}\widehat{\varphi}(\gamma)$$

for some $\{l^j_\lambda\}_{\lambda\in\Lambda} \in \ell^2(\mathbb{N}_0)$. By taking inverse Fourier transform on both sides, we obtain

$$\widehat{f}((\mathfrak{p}^{-1}N)^j\gamma) = \sum_{\lambda\in\Lambda} p^j_\lambda T_\lambda \varphi,$$

where

$$\sum_{\lambda\in\Lambda} |p^j_\lambda|^2 < \infty$$

which shows $f(\gamma) \in V_j$. Hence $V_j (j \in \mathbb{Z})$ are given by (4.4).

Now we are ready to show that $V_0 \subseteq V_1$. Let $f(\gamma) \in V_0$. Then by (4.4), we can write

$$f(\gamma) = \left\{ m^1_0(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} m^2_0(\gamma)\widehat{\varphi}(\gamma)\right\},$$

where m^1_0 and m^2_0 are locally integrable, q-periodic functions. Therefore,

$$\widehat{f}(\gamma) = G(\gamma)m(\gamma)\widehat{\varphi}(\gamma), \tag{4.5}$$

where

$$G(\gamma) = m^1_0(\mathfrak{p}^{-1}N\gamma) + \overline{\chi_{u(r)}(\mathfrak{p}^{-1}\gamma)} m^2_0(\mathfrak{p}^{-1}N\gamma)$$

and

$$m(\gamma) = m^1(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} m^2(\gamma).$$

This gives

$$G(\gamma)m(\gamma) = G(\gamma)\left\{ m^1(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} m^2(\gamma)\right\} = G(\gamma)m^1(\gamma) + \overline{\chi\left(\frac{u(r)}{N}\gamma\right)} G(\gamma)m^2(\gamma). \tag{4.6}$$

Using the conditions (i) and (iii), it can be easily shown that functions $m^1(\gamma)$ and $m^2(\gamma)$ are bounded. Also since $m^1(\gamma)$, $m^2(\gamma)$ and $G(\gamma)$ are q-periodic, therefore the functions $G(\gamma)m^1(\gamma)$ and $G(\gamma)m^2(\gamma)$ are q-periodic and

$$\int_{\mathfrak{D}} |G(\gamma)m^1(\gamma)|^2 d\gamma, \quad \int_{\mathfrak{D}} |G(\gamma)m^2(\gamma)|^2 d\gamma < \infty.$$

Thus by using (4.4)–(4.6), we infer that $f(\gamma) \in V_1$. Hence $V_0 \subseteq V_1$.

To prove that $\overline{\bigcup_{j\in\mathbb{Z}} V_0} = L^2(\mathbb{K})$, it sufficient to show that, for any $f \in L^2(\mathbb{K})$, we have

$$\|P_j f - f\|^2 = \|f\|^2 - \|P4.12_j f\|^2 \to 0 \quad \text{as} \quad j \to \infty,$$

where P_j is the orthonormal projection onto V_j. Let $f \in L^2(\mathbb{K})$ be such that $\widehat{f} \in C_c(\mathbb{K})$. Now we have

$$\|P_j f\|^2 = \sum_{\lambda\in\Lambda} |\langle f, D_{\mathfrak{p}^j} T_\lambda \varphi\rangle|^2 = \sum_{\lambda\in\Lambda} |\langle f, \widehat{D_{\mathfrak{p}^j} T_\lambda \varphi}\rangle|^2$$

$$\begin{aligned} &= \sum_{k\in\mathbb{N}_0} \left| \int_{\mathbb{K}} (qN)^{-1/2} \widehat{f}(\gamma) \chi_{u(k)}\left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\left(\frac{u(r)}{N} + \mathfrak{p}u(k)\right)\right) \overline{\widehat{\varphi}\left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\right)} d\gamma \right|^2 \\ &+ \sum_{k\in\mathbb{N}_0} \left| \int_{\mathbb{K}} (qN)^{-1/2} \widehat{f}(\gamma) \chi_{u(k)}\left(\frac{\gamma}{(\mathfrak{p}^{-1}N)^j}\left(\frac{u(r)}{N} + \mathfrak{p}u(k)\right)\right) \overline{\widehat{\varphi}\left(\frac{\gamma}{(2N)^j}\right)} d\gamma \right|^2 \end{aligned} \tag{4.7}$$

$$\begin{aligned} &= \sum_{k\in\mathbb{N}_0} \left| \int_{\mathbb{K}} (qN)^{j/2} \widehat{f}((\mathfrak{p}^{-1}N)^j\xi) \chi_{u(k)}(\mathfrak{p}^{-1}\xi) \overline{\widehat{\varphi}(\xi)} d\xi \right|^2 \\ &+ \sum_{k\in\mathbb{N}_0} \left| \int_{\mathbb{K}} (qN)^{j/2} \widehat{f}((\mathfrak{p}^{-1}N)^j\xi) \chi_{u(k)}(\mathfrak{p}^{-1}\xi) \overline{\widehat{\varphi}(\xi)} d\xi \right|^2. \end{aligned}$$

Since $\widehat{f}$ has compact support, we can choose j so large that

$$\operatorname{supp} \widehat{f}((\mathfrak{p}^{-1}N)^j\xi) \subseteq q^2\mathfrak{D}.$$

Then, using the fact that $\{\sqrt{q}\chi_{u(k)}(\xi)\}$ is an orthonormal basis for $L^2(q^2\mathfrak{D})$ and by (4.7), we get

$$\begin{aligned}\|P_j f\|^2 &= \frac{(qN)^j}{2}\Bigg\{\sum_{k\in\mathbb{N}_0}\left|\int_{q^2\mathfrak{D}}\widehat{f}((\mathfrak{p}^{-1}N)^j\xi)\sqrt{q}\chi_{u(k)}(\mathfrak{p}^{-1}\xi)\overline{\widehat{\varphi}(\xi)}d\xi\right|^2\\ &+\sum_{k\in\mathbb{N}_0}\left|\int_{q^2\mathfrak{D}}\widehat{f}((\mathfrak{p}^{-1}N)^j\xi)\sqrt{q}e^{2\pi i\frac{r}{N}\xi}\chi_{u(k)}(\mathfrak{p}^{-1}\eta)\overline{\widehat{\varphi}(\eta)}d\xi\right|^2\Bigg\}\\ &= (qN)^j\int_{q^2\mathfrak{D}}|\widehat{f}((\mathfrak{p}^{-1}N)^j\xi)\overline{\widehat{\varphi}(\xi)}|^2d\xi\end{aligned} \tag{4.8}$$

Putting $(\mathfrak{p}^{-1}N)^j\xi=\eta$ in (4.8) and invoking the Lesbesgue-dominated convergence theorem, we get

$$\|P_j f\|^2=\int_{(\mathfrak{p}^{-1}N)^{-j}\mathfrak{D}}\left|\widehat{f}(\eta)\overline{\widehat{\varphi}(\mathfrak{p}^{-1}N)^{-j}\eta}\right|^2d\eta\to\|f\|^2 \quad \text{as} \quad j\to\infty.$$

Thus the proof is complete. □

In the context of Fourier domain, the following theorem gives necessary condition for scaling function of wavelet NUMRA on LFPC.

Theorem 2. *If φ be a scaling function of wavelet NUMRA and $\widehat{\varphi}$ is continuous then $|\widehat{\varphi}(0)|=1$ and $\widehat{\varphi}\,(u(m)N-u(j))=0$, where $m\in\mathbb{N}_0,\ 0\le j\le N-1$. In particular $\widehat{\varphi}(u(m)N)=0$ for $m\in\mathbb{N}_0$ and $\widehat{\varphi}(-\mathfrak{p}u(j))=0,\quad 0\le j\le N-1$.*

P r o o f. By (4.3), we have

$$\lim_{j\to\infty}\int_{q^2\mathfrak{D}}|\widehat{\varphi}(\mathfrak{p}^{-1}N)^{-j}\mu|^2d\mu=\frac{1}{q}$$

as $|\widehat{\varphi}|$ is continuous. By virtue of Lebesgue dominated convergence theorem, we obtain $|\widehat{\varphi}(0)|=1$. Since φ is a scaling function for wavelet NUMRA, we have

$$\sum_{\gamma\in\Delta_N}|\widehat{\varphi}(\xi-\gamma)|^2=1 \quad \text{a.e.} \quad \xi\in\mathbb{K}. \tag{4.9}$$

Suppose

$$\widehat{\varphi}\,(u(m)N-\mathfrak{p}u(j))=a\neq 0$$

for some m,j not both zero together. Then

$$|\widehat{\varphi}(\xi)|+|\widehat{\varphi}\,(\xi+u(m)N-\mathfrak{p}u(j))|^2>1+a^2, \quad \text{when} \quad \xi\in\mathfrak{p}^{\epsilon}\mathfrak{D}$$

for some $\epsilon>0$ which contradicts (4.9). □

The following theorem gives the sufficient conditions for the frequency band of the scaling function of wavelet NUMRA on LFPC.

Theorem 3. *Let $\mho$ be a compact subset of $\mathbb{K}$ such that*

(i) $\mho \subseteq (\mathfrak{p}^{-1}N)\mho$;

(ii) $\bigcup_{m\in\mathbb{N}_0}(\mathfrak{p}^{-1}N)^j\mho = \mathbb{K}$;

(iii) $\sum_{j=0}^{N-1}\delta_{j/2}\star\sum_{m\in\mathbb{N}_0}\delta_{mN}\star\mathbf{\Phi}_\mho = 1.$

Then $\mho$ is the frequency band function for some wavelet NUMRA.

P r o o f. Let

$$V_j = \left\{f\in L^2(\mathbb{K}) : \operatorname{supp}\widehat{f}\subset(\mathfrak{p}^{-1}N)^j\mho, \quad j\in\mathbb{Z}\right\}$$

and $\psi\in L^2(\mathbb{K})$ be such that $\widehat{\varphi} = \mathbf{\Phi}_\mho$. Using hypothesis (i) and the definition of V_j, we have $V_j\subseteq V_{j+1}$ and $f((\mathfrak{p}^{-1}N)^j\gamma)\in V_j$ if and only if $f((\mathfrak{p}^{-1}N)^{j+1}\gamma)\in V_{j+1}$. By hypothesis (ii) and the definition of V_j, we get $\bigcup_{j\in\mathbb{Z}}V_j = L^2(\mathbb{K})$. By using Lemma 3 and hypothesis (iii), we get that $(\mho,\Lambda)$ is a spectral pair. Now we have

$$\widehat{T_\lambda\varphi}(\xi) = \overline{\chi_\lambda(\xi)}\widehat{\varphi}(\xi) = \overline{\chi_\lambda(\xi)}\mathbf{\Phi}_\mho(\xi)$$

and the Fourier transform is the unitary operator. Thus $\{T_\lambda\varphi\}_{\lambda\in\Lambda}$ is an orthonormal basis for V_0. By virtue of Lemma 3, we infer that $\bigcap_{j\in\mathbb{Z}}V_j = \{0\}$. Hence $\mho$ is frequency band for wavelet NUMRA (V_j,φ). □

5. Conclusion

In the present paper, we have given a complete characterization of the scaling function for the non-uniform multiresolution analysis on local fields of positive characteristic. Theorem 1 characterizes the nonzero square integrable functions on $L^2(\mathbb{K})$ to be a scaling functions for the wavelet NUMRA by means of three simple conditions. Furthermore Theorem 3 expresses a compact subset of $\mathbb{K}$ to be the band scaling function of wavelet NUMRA on LFPC by means of three conditions. The present study can be extended in fractional settings and in the context of Multiresolution Analysis associated with Linear Canonical Transform.

Acknowledgements

The authors pay gratitude to the referees for their valuable suggestions and comments.

REFERENCES

1. Ahmad I., Sheikh N. A. a-inner product on local fields of positive characteristic. *J. Nonlinear Anal. Appl.*, 2018. Vol. 2018, No. 2. P. 64–75.
2. Ahmad I., Sheikh N. A. Dual wavelet frames in Sobolev spaces on local fields of positive characteristic. *Filomat*, 2020. Vol. 34, No. 6. P. 2091–2099. DOI: 10.2298/FIL2006091A
3. Ahmad O., Sheikh N. A. Explicit construction of tight nonuniform framelet packets on local fields. *Oper. Matrices*, 2021. Vol. 15, No. 1. P. 131–149. DOI: 10.7153/oam-2021-15-10
4. Ahmad O., Sheikh N. A. On characterization of nonuniform tight wavelet frames on local fields. *Anal. Theory Appl.*, 2018. Vol. 34. P. 135–146. DOI: 10.4208/ata.2018.v34.n2.4
5. Albeverio S., Evdokimov S., Skopina M. p-adic multiresolution analysis and wavelet frames. *J. Fourier Anal. Appl.*, 2010. Vol. 16. P. 693–714. DOI: 10.1007/s00041-009-9118-5
6. Albeverio S., Kozyrev S. Multidimensional basis of p-adic wavelets and representation theory. *p-Adic Num. Ultrametric Anal. Appl.*, 2009. Vol. 1. No. 3. P. 181–189. DOI: 10.1134/S2070046609030017

7. Behera B., Jahan Q. Multiresolution analysis on local fields and characterization of scaling functions. *Adv. Pure Appl. Math.*, 2012. Vol. 3, No. 2. P. 181–202. DOI: 10.1515/apam-2011-0016
8. Behera B., Jahan Q. Characterization of wavelets and MRA wavelets on local fields of positive characteristic. *Collect. Math.*, 2015. Vol. 66, No. 1. P. 33–53. DOI: 10.1007/s13348-014-0116-9
9. Benedetto J. J., Benedetto R. L. A wavelet theory for local fields and related groups. *J. Geom. Anal.*, 2004. Vol. 14, No. 3. P. 423–456.
10. Cifuentes P., Kazarian K. S., Antolín A. S. Characterization of scaling functions in multiresolution analysis. *Proc. Am. Math. Soc.*, 2005. Vol. 133, No. 4. P. 1013–1023.
11. Gabardo J.-P., Nashed M.Z. Nonuniform multiresolution analyses and spectral pairs. *J. Funct. Anal.*, 1998. Vol. 158, No. 1. P. 209–241. DOI: 10.1006/jfan.1998.3253
12. Gabardo J.-P., Yu X. Wavelets associated with nonuniform multiresolution analyses and one-dimensional spectral pairs. *J. Math. Anal. Appl.*, 2006. Vol. 323, No. 2. P. 798–817. DOI: 10.1016/j.jmaa.2005.10.077
13. Jiang H., Li D., Jin N. Multiresolution analysis on local fields. *J. Math. Anal. Appl.*, 2004. Vol. 294, No. 2. P. 523–532. DOI: 10.1016/j.jmaa.2004.02.026
14. Khrennikov A. Yu., Kozyrev S. V. Wavelets on ultrametric spaces. *Appl. Comput. Harmon. Anal.*, 2005. Vol. 19. P. 61–76. DOI: 10.1016/j.acha.2005.02.001
15. Khrennikov A. Yu., Shelkovich V. M. An infinite family of p-adic non-Haar wavelet bases and pseudo-differential operators. *p-Adic Num. Ultrametric Anal. Appl.*, 2009. Vol. 1. P. 204–216. DOI: 10.1134/S2070046609030030
16. Khrennikov A. Yu., Shelkovich V. M. Skopina M. p-adic orthogonal wavelet bases. *p-Adic Num. Ultrametric Anal. Appl.*, 2009. Vol. 1, No. 2. P. 145–156. DOI: 10.1134/S207004660902006X
17. Khrennikov A. Yu., Shelkovich V. M. Skopina M. p-adic refinable functions and MRA-based wavelets. *J. Approx. Theory*, 2009. Vol. 161, No. 1. P. 226–238. DOI: 10.1016/j.jat.2008.08.008
18. Kozyrev S. V. Wavelet theory as p-adic spectral analysis. *Izv. Math.*, 2002. Vol. 66, No. 2. P. 149–158.
19. Li D., Jiang H. The necessary condition and sufficient conditions for wavelet frame on local fields. *J. Math. Anal. Appl.*, 2008. Vol. 345, No. 1. P. 500–510. DOI: 10.1016/j.jmaa.2008.04.031
20. Madych W.R. Some elementary properties of multiresolution analysis of $L^2(\mathbb{R}^n)$. In: *Wavelets: A Tutorial in Theory and Applications. Vol. 2: Wavelet Analysis and Its Applications. Chui C. K. (ed.)*, 1992. P. 259–294. DOI: 10.1016/B978-0-12-174590-5.50015-0
21. Mallat S. G. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 1989. Vol. 315, No. 1. P. 69–87.
22. Shah F. A., Ahmad O. Wave packet systems on local fields. *J. Geom. Phys.*, 2017. Vol. 120. P. 5–18. DOI: 10.1016/j.geomphys.2017.05.015
23. Shah F. A., Abdullah. Nonuniform multiresolution analysis on local fields of positive characteristic. *Complex Anal. Oper. Theory*, 2015. Vol. 9. P. 1589–1608. DOI: 10.1007/s11785-014-0412-0
24. Shukla N. K., Maury S. C. Super-wavelets on local fields of positive characteristic. *Math. Nachr.*, 2018. Vol. 291, No. 4. P. 704–719. DOI: 10.1002/mana.201500344
25. Taibleson M. H. *Fourier Analysis on Local Fields. (MN-15).* Princeton, NJ: Princeton University Press, 1975. 306 p.
26. Zhang Z. Supports of Fourier transforms of scaling functions. *Appl. Comput. Harmon. Anal.*, 2007. Vol. 22, No. 2. P. 141–156. DOI: 10.1016/j.acha.2006.05.007

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 16–25
DOI: 10.15826/umj.2021.1.002

SOME REMARKS ON ROUGH STATISTICAL Λ-CONVERGENCE OF ORDER α

Reena Antal

Department of Mathematics, Chandigarh University,
NH-95, Chandigarh-Ludhiana Highway, Mohali, Punjab 140413, India
reena.antal@gmail.com

Meenakshi Chawla

Department of Mathematics, Chandigarh University,
NH-95, Chandigarh-Ludhiana Highway, Mohali, Punjab 140413, India
chawlameenakshi7@gmail.com

Vijay Kumar

Department of Mathematics, Panipat Institute of Engineering and Technology,
70, Milestone GT Road, Samalkha, Panipat, 132102 Haryana, India
vjy_kaushik@yahoo.com

Abstract: The main purpose of this work is to define Rough Statistical Λ-Convergence of order α $(0 < \alpha \leq 1)$ in normed linear spaces. We have proved some basic properties and also provided some examples to show that this method of convergence is more generalized than the rough statistical convergence. Further, we have shown the results related to statistically Λ-bounded sets of order α and sets of rough statistically Λ-convergent sequences of order α.

Keywords: Statistical convergence, Rough statistical convergence, Rough statistical limit points.

1. Introduction

In 1951, Fast [5] presented a new idea of convergence named as statistical convergence that is more generalized than the usual convergence for the sequences.

Definition 1 [5]. *A sequence $x = \{x_m\}$ of numbers is said to be statistically convergent to ξ if for every $\epsilon > 0$ we have $\lim\limits_{n\to\infty} |M(x,\epsilon)|/n = 0$, where $|M(x,\epsilon)|$ represents the order of the enclosed set $M(x,\epsilon) = \{m \leq n : |x_m - \xi| \geq \epsilon\}$.*

This idea has interesting applications in the field of Fourier Analysis [1], Measure Theory [16], Approximation Theory [7] etc. It has been studied by many researchers for various types of sequences in different setups like locally convex spaces [10], probabilistic normed spaces [8], random normed spaces [3], intuitionistic fuzzy normed spaces [9] etc.

An interesting generalization of usual convergence named as rough convergence was introduced by Phu [19] for the sequences in finite dimensional normed linear spaces and later on introduced on infinite dimensional normed linear spaces [20]. He mainly worked on rough limits, roughness degree, rough continuity of linear operators and also introduced rough Cauchy sequences.

Definition 2 [19]. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is said to be rough convergent to $\xi \in \mathbb{X}$ if for every $\epsilon > 0$ there exists a non-negative number r and $m_0 \in \mathbb{N}$ such that $\|x_m - \xi\| < r + \epsilon$, for all $m \geq m_0$.*

Aytar [2] extended the rough convergence to rough statistical convergence like usual convergence is extended to statistical convergence with the help of natural density.

Definition 3 [2]. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is said to be rough statistically convergent to $\xi \in \mathbb{X}$ if for every $\epsilon > 0$ there exists a non-negative number r such that*

$$\lim_{n\to\infty} \frac{1}{n}\left|\{m \leq n : \|x_m - \xi\| \geq r + \epsilon\}\right| = 0,$$

where ξ is known as r-St-limit of sequence $x = \{x_m\}$.

Aytar [2] also defined the rough statistical bounded sequence along with the set of rough statistical limit points of a sequence. Further, some criterion associated with the convexity and closeness of the set of rough statistical limit points of a sequence was investigated.

Inspired by the work of Aytar [2], Maity [12] presented the concept of rough statistical convergence of order α $(0 < \alpha \leq 1)$ in normed linear spaces and explained some important results for the set of rough statistical limit points of order α. The idea of pointwise rough statistical convergence and rough statistical Cauchy sequences for real valued functions was introduced in [11]. The concept of rough convergence has been defined for double sequences by Malik and Maity in [13] and after that the authors extended this idea in [14] and defined rough statistical convergence for double sequences in normed linear spaces.

This idea has motivated many authors to use the concepts of ideals also. Pal et al. [18] introduced rough I-convergence with the help of ideals of $\mathbb{N}$. Later, Malik et al. in [15] extended this concept of rough I-convergence to rough I-statistical convergence and described some topological properties of the set of all rough I-statistical limits of sequences in normed linear spaces. A lot of work has been done on rough convergence and its generalizations. More investigations and applications of rough convergence can be revealed as it is an active area of research.

In this paper, we are introducing the concept of rough statistical Λ-convergence of order α $(0 < \alpha \leq 1)$ in the normed linear spaces.

2. Main results

In order to study the basic concept of rough statistical Λ-convergence, we first consider a sequence $\lambda = \{\lambda_j\}$ of real numbers such that $0 < \lambda_0 < \lambda_1 < ...... < \lambda_j < ...$ and $\lambda_j \to \infty$ as $j \to \infty$. The concept of Λ-convergence for real sequences have been defined by Mursaleen[17] as given below: a sequence $x = \{x_m\}$ of real numbers is Λ-convergent to a number L if $\Lambda x_m \to L$ as $m \to \infty$ where

$$\Lambda x_m = \frac{1}{\lambda_m} \sum_{j=0}^{m} (\lambda_j - \lambda_{j-1}) x_j.$$

Here, without loss of generality we take all the terms with negative subscripts equal to zero.

Using this concept, we are defining the notion of the rough Λ-convergence and rough statistical Λ-convergence as follows:

Definition 4. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is said to be rough Λ-convergent to $\xi \in \mathbb{X}$ if for every $\epsilon > 0$ there exist a non-negative number r and $m_0 \in \mathbb{N}$ such that $\|\Lambda x_m - \xi\| < r + \epsilon$, for all $m \geq m_0$.*

Definition 5. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is said to be rough statistically Λ-convergent to ξ if for every $\epsilon > 0$ there exists some non-negative number r such that*

$$\lim_{n\to\infty} \frac{1}{n}\big|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}\big| = 0,$$

where ξ is known as r-St_Λ-limit of sequence $x = \{x_m\}$.

Remark 1. For the case $r = 0$, the notion of rough statistical Λ-convergence agrees with the statistical Λ-convergence.

Çolak [4] has given an interesting idea related to the statistical convergence of order α $(0 < \alpha \leq 1)$ with the help of α-density. Motivated by his idea, now we are defining a rough statistical Λ-convergence of order α $(0 < \alpha \leq 1)$ as follows:

Definition 6. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is said to be rough statistically Λ-convergent of order α $(0 < \alpha \leq 1)$ to the number $\xi \in \mathbb{X}$ if for every $\epsilon > 0$ there exists some non-negative number r such that*

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}\big| = 0,$$

where ξ is known as r-St^α_Λ-limit of sequence $x = \{x_m\}$. It is denoted by

$$x_m \xrightarrow{r\text{-}St^\alpha_\Lambda} \xi.$$

The set of all the rough statistically Λ-convergent sequences of order $\alpha(0 < \alpha \leq 1)$ is denoted by rSt^α_Λ for fixed r.

In general, the r-St^α_Λ-limit of a sequence may be not unique. So we consider r-St^α_Λ-limit set of a sequence $x = \{x_m\}$ as

$$r\text{-}St^\alpha_\Lambda\text{-}LT_x = \{\xi : x_m \xrightarrow{r\text{-}St^\alpha_\Lambda} \xi\}.$$

The sequence $x = \{x_m\}$ is said to be r-St^α_Λ-convergent such that r-St^α_Λ-$LT_x \neq \phi$. For unbounded sequence the rough limit set is always empty.
But in case of rough statistical Λ-convergence of order α, we have r-St^α_Λ-$LT_x \neq \phi$ even though sequence may be unbounded. For this we have given the next example.

Example 1. Let $\mathbb{X} = \mathbb{R}$. Then, define a sequence

$$\Lambda x_m = \begin{cases} (-1)^m, & m \neq n^2, \\ m, & \text{otherwise}. \end{cases}$$

Take $\alpha = 1$, then

$$r\text{-}St^\alpha_\Lambda\text{-}LT_x = \begin{cases} \phi, & r < 1, \\ [1 - r, r - 1], & \text{otherwise} \end{cases}$$

and r-Λ-$LT_x = \phi$ for all $r \geq 0$. Thus, this sequence is divergent in ordinary sense as it is unbounded. Also, the sequence is not statistically Λ-convergent for any r.

With the help of statistically cluster points defined by Fridy [6], we are giving the following definition as follows:

Definition 7. *A point ξ is said to be rough statistically Λ-cluster point of order α $(0 < \alpha \leq 1)$ of a sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ if for every $\epsilon > 0$ there exists some non-negative number r such that*

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}\big| \neq 0.$$

Definition 8. *A sequence $x = \{x_m\}$ is said to be statistically Λ-bounded if there exists a real number $M_0 > 0$ such that*

$$\lim_{n\to\infty} \frac{1}{n}\big|\{m \le n : \|\Lambda x_m\| \ge M_0\}\big| = 0.$$

Definition 9. *A sequence $x = \{x_m\}$ is said to be statistically Λ-bounded of order α $(0 < \alpha \le 1)$ if there exists a real number $M_0 > 0$ such that*

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \le n : \|\Lambda x_m\| \ge M_0\}\big| = 0.$$

In view of above definitions, we obtained the following interesting results on rough statistical Λ-convergence.

Theorem 1. *Every rough Λ-convergent sequence is also rough statistically Λ-convergent of order α $(0 < \alpha \le 1)$, but converse may be not true.*

P r o o f. Let the sequence $x = \{x_m\}$ be rough Λ-convergent in a normed linear space $(\mathbb{X}, \|\cdot\|)$. Then, for every $\epsilon > 0$ and some $r > 0$ there exists a real number $M_0 > 0$ such that $\|\Lambda x_m - \xi\| \ge r + \epsilon$ for all $m \ge M_0$.

The set $\{m \le n : \|\Lambda x_m - \xi\| \ge r + \epsilon\}$ has finitely many terms. Thus,

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \le n : \|\Lambda x_m - \xi\| \ge r + \epsilon\}\big| = 0.$$

Hence, the sequence $x = \{x_m\}$ is rough statistically Λ-convergent of order α $(0 < \alpha \le 1)$.

But the contrary part is not true which can be justified by the next example.

Example 2. Consider the normed space $(\mathbb{R}, \|\cdot\|)$ under the usual norm. Define a sequence

$$\Lambda x_m = \begin{cases} 1, & m \text{ is a square,} \\ 0, & \text{otherwise.} \end{cases}$$

For $\epsilon > 0$ and some $r \ge 0$ we have

$$\begin{aligned} M(r,\epsilon) &= \{m \le n : \|\Lambda x_m - \xi\| \ge r + \epsilon\}; \quad \xi = 0 \\ &= \{m \le n : \|\Lambda x_m\| \ge r + \epsilon > 0\} \\ &= \{m \le n : \|\Lambda x_m\| = 1\} \\ &= \{m \le n : m \text{ is a square}\}. \end{aligned}$$

Thus,

$$\lim_{n\to\infty} \frac{1}{n^\alpha}|M(r,\epsilon)| \le \lim_{n\to\infty} \frac{\sqrt{n}}{n^\alpha} = 0.$$

Therefore, $x = \{x_m\}$ is rough statistically Λ-convergent of order α to 0 for $\alpha > 1/2$.

□

In the next theorem we discuss the algebraic characterization of rough statistically Λ-convergent sequences of order α $(0 < \alpha \le 1)$.

Theorem 2. *Let $x = \{x_m\}$ and $y = \{y_m\}$ be two sequences in a normed linear space $(\mathbb{X}, \|\cdot\|)$ and α $(0 < \alpha \le 1)$ be given. Then for some non-negative number r the following holds:*

(1) *if* $x_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} x_0$ *and* $k \in \mathbb{N}$ *then* $kx_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} kx_0$;

(2) *if* $x_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} x_0$ *and* $y_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} y_0$ *then* $(x_m + y_m) \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} (x_0 + y_0)$.

P r o o f. (1) If $k = 0$ then there is nothing to prove.

If $k \neq 0$. Since $x_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} x_0$ then for given $\epsilon > 0$ and some $r \geq 0$, we have the set

$$M(r,\epsilon) = \Big\{m \leq n : \|\Lambda x_m - x_0\| \geq \frac{r+\epsilon}{|k|}\Big\} \quad \text{with} \quad \lim_{n\to\infty} \frac{1}{n^{\alpha}} |M(r,\epsilon)| = 0.$$

Let $m \in M^c(r,\epsilon)$. Then

$$\|\Lambda k x_m - kx_0\| = |k|\, \|\Lambda x_m - x_0\| < |k| \Big(\frac{r+\epsilon}{|k|}\Big) < r + \epsilon.$$

This implies that

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \Big|\Big\{m \leq n : \|\Lambda k x_m - kx_0\| < \frac{r+\epsilon}{|k|}\Big\}\Big| = 1,$$

i. e.

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \Big|\Big\{m \leq n : \|\Lambda k x_m - kx_0\| \geq \frac{r+\epsilon}{|k|}\Big\}\Big| = 0.$$

Therefore, $kx_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} kx_0$.

(2) Since $x_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} x_0$ and $y_m \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} y_0$ then for given $\epsilon > 0$ and some $r \geq 0$, we have sets

$$M_x(r,\epsilon) = \Big\{m \leq n : \|\Lambda x_m - x_0\| \geq \frac{r+\epsilon}{2}\Big\} \quad \text{with} \quad \lim_{n\to\infty} \frac{1}{n^{\alpha}} |M_x(r,\epsilon)| = 0,$$

$$M_y(r,\epsilon) = \Big\{m \leq n : \|\Lambda y_m - y_0\| \geq \frac{r+\epsilon}{2}\Big\} \quad \text{with} \quad \lim_{n\to\infty} \frac{1}{n^{\alpha}} |M_y(r,\epsilon)| = 0.$$

Let $m \in M^c_x(r,\epsilon) \cap M^c_y(r,\epsilon)$. Then

$$\|\Lambda(x_m + y_m) - (x_0 + y_0)\| \leq \|\Lambda x_m - x_0\| + \|\Lambda y_m - y_0\| < \frac{r+\epsilon}{2} + \frac{r+\epsilon}{2} = r + \epsilon.$$

This implies that

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \big|\{m \leq n : \|\Lambda(x_m + y_m) - (x_0 + y_0)\| < r + \epsilon\}\big| = 1,$$

i. e.

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \big|\{m \leq n : \|\Lambda(x_m + y_m) - (x_0 + y_0)\| \geq r + \epsilon\}\big| = 0.$$

Therefore, $(x_m + y_m) \xrightarrow{r\text{-}St^{\alpha}_{\Lambda}} (x_0 + y_0)$. □

Theorem 3. *Let* $0 < \alpha \leq \beta \leq 1$ *then* $rSt^{\alpha}_{\Lambda} \subseteq rSt^{\beta}_{\Lambda}$ *where* rSt^{α}_{Λ} *and* rSt^{β}_{Λ} *represent the sets of all rough statistically* Λ*-convergent of order* α *and* β *respectively.*

P r o o f. Let $x = \{x_m\}$ be a sequence in a normed linear space $(\mathbb{X}, \|\cdot\|)$. If $0 < \alpha \leq \beta \leq 1$ then for every $\epsilon > 0$ and some $r > 0$ with the limit point ξ, we have

$$\frac{1}{n^{\beta}} \big|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}\big| \leq \frac{1}{n^{\alpha}} \big|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}\big|.$$

Therefore, we get $rSt^{\alpha}_{\Lambda} \subseteq rSt^{\beta}_{\Lambda}$. □

Theorem 4. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is statistically Λ-bounded of order $\alpha(0 < \alpha \leq 1)$ if and only if r-St^{α}_{Λ}-$LT_x \neq \phi$, for some non-negative number r.*

P r o o f. Let the sequence $x = \{x_m\}$ is statistically Λ-bounded of order α $(0 < \alpha \leq 1)$, then there exists a real number $M_0 > 0$ such that

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \big|\{m \leq n : \|\Lambda x_m\| \geq M_0\}\big| = 0.$$

Let $M = \big\{m \in \mathbb{N} : \|\Lambda x_m\| \geq M_0\big\}$. Define $r_0 = \sup\big\{\|\Lambda x_m\| : m \in M^c\big\}$. As

$$0 \in r_0\text{-}St^{\alpha}_{\Lambda}\text{-}LT_x \Rightarrow r_0\text{-}St^{\alpha}_{\Lambda}\text{-}LT_x \neq \phi.$$

Conversely, suppose that r-St^{α}_{Λ}-$LT_x \neq \phi$ for some $r \geq 0$. Then, for each $\epsilon > 0$ there exists $\xi \in \mathbb{X}$ such that $\xi \in r$-St^{α}_{Λ}-LT_x. Then

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \big|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}\big| = 0.$$

Hence, the sequence $x = \{x_m\}$ is statistically Λ-bounded of order α. □

Theorem 5. *If $x' = \{x_{m_k}\}$ is a non-thin subsequence of a sequence $x = \{x_m\}$ then*

$$r\text{-}St^{\alpha}_{\Lambda}\text{-}LT_x \subseteq r\text{-}St^{\alpha}_{\Lambda}\text{-}LT_{x'}.$$

P r o o f. The proof of above results is obvious, so we are omitting it. □

Theorem 6. *Let $x = \{x_m\}$ be a sequence in a normed linear space $(\mathbb{X}, \|\cdot\|)$. Then, the rough statistical limit set of order α $(0 < \alpha \leq 1)$ is convex, i.e., r-St^{α}_{Λ}-LT_x is convex.*

P r o o f. Let $\xi_1, \xi_2 \in r$-St^{α}_{Λ}-LT_x and $\epsilon > 0$ be given. For the convexity of the set r-St^{α}_{Λ}-LT_x, we have to show that $[(1-\beta)\xi_1 + \beta\xi_2] \in r$-$St^{\alpha}_{\Lambda}$-$LT_x$ for some $\beta \in (0,1)$. Now, we define

$$M_1(r,\epsilon) = \Big\{m \in \mathbb{N} : \|\Lambda x_m - \xi_1\| \geq \frac{r+\epsilon}{2(1-\beta)}\Big\},$$

$$M_2(r,\epsilon) = \Big\{m \in \mathbb{N} : \|\Lambda x_m - \xi_2\| \geq \frac{r+\epsilon}{2\beta}\Big\}.$$

As $\xi_1, \xi_2 \in r$-St^{α}_{Λ}-LT_x, we have

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} |M_1(r,\epsilon)| = \lim_{n\to\infty} \frac{1}{n^{\alpha}} |M_2(r,\epsilon)| = 0.$$

Let $m \in M_1^c(r,\epsilon) \cap M_2^c(r,\epsilon)$. Then

$$\begin{aligned}\big\|\Lambda x_m - [(1-\beta)\xi_1 + \beta\xi_2]\big\| &= \big\|(1-\beta)(\Lambda x_m - \xi_1) + \beta(\Lambda x_m - \xi_2)\big\| \\ &\leq (1-\beta)\big\|\Lambda x_m - \xi_1\big\| + \beta\big\|\Lambda x_m - \xi_2\big\| \\ &< r + \epsilon.\end{aligned}$$

Since

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} |M_1^c(r,\epsilon) \cap M_2^c(r,\epsilon)| = 1,$$

we get

$$\lim_{n\to\infty} \frac{1}{n^{\alpha}} \big|\{m \leq n : \|\Lambda x_m - [(1-\beta)\xi_1 + \beta\xi_2]\| \geq r + \epsilon\}\big| = 0,$$

i. e.

$$[(1-\beta)\xi_1 + \beta\xi_2] \in r\text{-}St^{\alpha}_{\Lambda}\text{-}LT_x.$$

Hence, r-St^{α}_{Λ}-LT_x is a convex set. □

Theorem 7. *A sequence $x = \{x_m\}$ in a normed linear space $(\mathbb{X}, \|\cdot\|)$ is rough statistically Λ-convergent of order α $(0 < \alpha \leq 1)$ to $\xi \in \mathbb{X}$ for some non-negative number r if and only if there exists a sequence $y = \{y_m\}$ in $\mathbb{X}$ which is rough statistically Λ-convergent of order α to ξ and $\|\Lambda x_m - \Lambda y_m\| \leq r$ for all $m \in \mathbb{N}$.*

P r o o f. *Necessity.* Let $x_m \xrightarrow{r\text{-}St_\Lambda^\alpha} \xi$. Then, for each $\epsilon > 0$ and some $r > 0$ we have

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\Big|\big\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\big\}\Big| = 0. \tag{2.1}$$

Now, we define the sequence as

$$\Lambda y_m = \begin{cases} \xi, & \|\Lambda x_m - \xi\| \leq r, \\ \Lambda x_m + r\dfrac{\xi - \Lambda x_m}{\|\Lambda x_m - \xi\|}, & \text{otherwise.} \end{cases}$$

Then, we have

$$\Lambda y_m - \xi = \begin{cases} 0, & \|\Lambda x_m - \xi\| \leq r, \\ \dfrac{\Lambda x_m - \xi}{\|\Lambda x_m - \xi\|}\big(\|\Lambda x_m - \xi\| - r\big), & \text{otherwise} \end{cases}$$

such that $\|\Lambda x_m - \Lambda y_m\| \leq r$ for all $m \in \mathbb{N}$. Further,

$$\|\Lambda y_m - \xi\| = \begin{cases} 0, & \|\Lambda x_m - \xi\| \leq r, \\ \|\Lambda x_m - \xi\| - r, & \text{otherwise.} \end{cases}$$

Hence, by the definition of Λy_m and (2.1), we have

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\Big|\big\{m \leq n : \|\Lambda y_m - \xi\| \geq r + \epsilon\big\}\Big| = 0,$$

which prove that the sequence $y = \{y_m\}$ is rough statistically Λ-convergent of order α to ξ.

Sufficiency. Since the sequence $y = \{y_m\}$ is rough statistically Λ-convergent of order α $(0 < \alpha \leq 1)$ to ξ then for $\epsilon > 0$ we have

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\Big|\big\{m \leq n : \|\Lambda y_m - \xi\| \geq r + \epsilon\big\}\Big| = 0.$$

Now for some $r > 0$ and sequence $x = \{x_m\}$ with $\|\Lambda x_m - \Lambda y_m\| \leq r$, the following inclusion holds

$$\big\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\big\} \subseteq \big\{m \leq n : \|\Lambda y_m - \xi\| \geq r + \epsilon\big\}.$$

Hence, we get

$$\lim_{n\to\infty} \frac{1}{n^\alpha}|\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\}| = 0.$$

□

Theorem 8. *The set $r\text{-}St_\Lambda^\alpha\text{-}LT_x$ of rough statistical Λ-limit set of order α $(0 < \alpha \leq 1)$ is closed.*

P r o o f. (i) If $r\text{-}St_{\Lambda}^{\alpha}\text{-}LT_x = \phi$, then we have to prove nothing.

(ii) If $r\text{-}St_{\Lambda}^{\alpha}\text{-}LT_x \neq \phi$. Then, take a sequence $y = \{y_m\} \subseteq r\text{-}St_{\Lambda}^{\alpha}\text{-}LT_x$ such that $\Lambda y_m \to y_*$ for $m \to \infty$. It is sufficient to show that $y_* \in r\text{-}St_{\Lambda}^{\alpha}\text{-}LT_x$.

As $\Lambda y_m \to y_*$, then for given $\epsilon > 0$ there exists $m_\epsilon \in \mathbb{N}$ such that

$$\|\Lambda y_m - y_*\| < \frac{r+\epsilon}{3}$$

for $m > m_\epsilon$.

Now choose $m_0 \in \mathbb{N}$ such that $m_0 > m_\epsilon$. Then we have

$$\|\Lambda y_{m_0} - y_*\| < \frac{r+\epsilon}{3}.$$

Again as $y = \{y_m\} \subseteq r\text{-}St_{\Lambda}^{\alpha}\text{-}LT_x$, we have $y_{m_0} \in r\text{-}St_{\Lambda}^{\alpha}\text{-}LT_x$. Clearly,

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\left|\left\{m \leq n : \|\Lambda x_m - y_{m_0}\| \geq \frac{r+\epsilon}{3}\right\}\right| = 0. \tag{2.2}$$

Next we prove the inclusion

$$\left\{m \leq n : \|\Lambda x_m - y_{m_0}\| < \frac{r+\epsilon}{3}\right\} \subseteq \{m \leq n : \|\Lambda x_m - y_*\| < r+\epsilon\}. \tag{2.3}$$

Let

$$k \in \left\{m \leq n : \|\Lambda x_m - y_{m_0}\| < \frac{r+\epsilon}{3}\right\} \Rightarrow \|\Lambda x_k - y_{m_0}\| < \frac{r+\epsilon}{3}.$$

Hence,

$$\|\Lambda x_k - y_*\| = \|\Lambda x_k - y_{m_0} + \Lambda y_m - y_* - \Lambda y_m + y_{m_0}\| \leq \|\Lambda x_k - y_{m_0}\| + \|\Lambda y_m - y_*\| + \|\Lambda y_k - y_{m_0}\|.$$

Using equation (2.2) and Theorem 7 we get

$$\|\Lambda y_k - y_{m_0}\| < \frac{r+\epsilon}{3}.$$

Thus,

$$\|\Lambda x_k - y_*\| < \frac{r+\epsilon}{3} + \frac{r+\epsilon}{3} + \frac{r+\epsilon}{3} = r+\epsilon.$$

This implies that

$$k \in \{m \leq n : \|\Lambda x_m - y_*\| < r+\epsilon\}.$$

Hence the inclusion (2.3) is proved.

Thus,

$$\{m \leq n : \|\Lambda x_m - y_*\| \geq r+\epsilon\} \subseteq \left\{m \leq n : \|\Lambda x_m - y_{m_0}\| \geq \frac{r+\epsilon}{3}\right\}.$$

Now,

$$\lim_{n\to\infty} \frac{1}{n^\alpha}|\{m \leq n : \|\Lambda x_m - y_*\| \geq r+\epsilon\}| \leq \lim_{n\to\infty} \frac{1}{n^\alpha}\left|\left\{m \leq n : \|\Lambda x_m - y_{m_0}\| \geq \frac{r+\epsilon}{3}\right\}\right|. \tag{2.4}$$

Using equation (2.2), we obtained that the set on left side of (2.4) has density 0. Hence, we get

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \leq n : \|\Lambda x_m - y_*\| \geq r+\epsilon\}\big| = 0.$$

□

Theorem 9. *Let $\Gamma_{\Lambda x}$ be the set of all rough statistical Λ-cluster points of order α $(0 < \alpha \leq 1)$ for a sequence $x = \{x_m\}$ in the normed linear space $(\mathbb{X}, \|\cdot\|)$. Then for an arbitrary $c \in \Gamma_{\Lambda x}$ and a positive real number r, we have $\|\xi - c\| < r$ for all $\xi \in r\text{-}St_\Lambda^\alpha\text{-}LT_x$.*

P r o o f. We prove the result by contradiction. For given α $(0 < \alpha \leq 1)$, we take a point $c \in \Gamma_{\Lambda x}$ and $\xi \in r\text{-}St_\Lambda^\alpha\text{-}LT_x$ such that $\|\xi - c\| > r$. By choosing $\epsilon = (\|\xi - c\| - r)/3$, we get the following inclusion

$$\big\{m \leq n : \|\Lambda x_m - \xi\| \geq r + \epsilon\big\} \supseteq \big\{m \leq n : \|\Lambda x_m - c\| < \epsilon\big\}. \tag{2.5}$$

Since $c \in \Gamma_{\Lambda x}$, then

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \leq n : \|\Lambda x_m - c\| < \epsilon\}\big| \neq 0.$$

By (2.5), we get

$$\lim_{n\to\infty} \frac{1}{n^\alpha}\big|\{m \leq n : \|\Lambda x_m - \xi\| < r + \epsilon\}\big| \neq 0,$$

which is a contradiction to $\xi \in r\text{-}St_\Lambda^\alpha\text{-}LT_x$. □

Theorem 10. *Let $x = \{x_m\}$ be a sequence in a strictly convex normed linear space $(\mathbb{X}, \|\cdot\|)$. Let α and r be two positive real numbers. If any $\xi_0, \xi_1 \in r\text{-}St_\Lambda^\alpha\text{-}LT_x$ with $\|\xi_0 - \xi_1\| = 2r$, then $x = \{x_m\}$ is rough statistically Λ-convergent of order α $(0 < \alpha \leq 1)$ to $(\xi_0 + \xi_1)/2$.*

P r o o f. Let $z \in \Gamma_{\Lambda x}$ and $\xi_0, \xi_1 \in r\text{-}St_\Lambda^\alpha\text{-}LT_x$ such that $\|\xi_0 - \xi_1\| = 2r$. Then, we have

$$\|\xi_0 - z\| \leq r \quad \text{and} \quad \|\xi_1 - z\| \leq r, \tag{2.6}$$

and by triangle inequality, we get

$$\begin{aligned} &\|\xi_0 - \xi_1\| \leq \|\xi_0 - z\| + \|\xi_1 - z\| \\ &\Rightarrow 2r \leq \|\xi_0 - z\| + \|\xi_1 - z\|. \end{aligned} \tag{2.7}$$

We get from (2.6) and (2.7)

$$\|\xi_0 - z\| = \|\xi_1 - z\| = r.$$

Also

$$\frac{1}{2}(\xi_1 - \xi_0) = \frac{1}{2}\big[(z - \xi_0) + (\xi_1 - z)\big], \tag{2.8}$$

and using $\|\xi_0 - \xi_1\| = 2r$, we get $(\xi_1 - \xi_0)/2 = r$.

Now from equation (2.8) and from strict convexity of the normed linear space $(\mathbb{X}, \|\cdot\|)$, we have $(z - \xi_0) = (\xi_1 - z) = (\xi_1 - \xi_0)/2$ which implies that $z = (\xi_0 + \xi_1)/2$. Thus, z is a unique statistical Λ-cluster point of sequence $x = \{x_m\}$.

As $\xi_0, \xi_1 \in r\text{-}St_\Lambda^\alpha\text{-}LT_x \Rightarrow r\text{-}St_\Lambda^\alpha\text{-}LT_x \neq \phi$. Hence, by Theorem 4, the sequence $x = \{x_m\}$ is statistically Λ-bounded of order α.

Since z is the unique statistical Λ-cluster point to statistically Λ-bounded sequence $x = \{x_m\}$ of order α.

This implies that $x_m \xrightarrow{r\text{-}St_\Lambda^\alpha} z$, where $z = (\xi_0 + \xi_1)/2$. □

Acknowledgements

We express great sense of gratitude and deep respect to the referees of this paper for their valuable suggestions.

REFERENCES

1. Alotaibi A., Mursaleen M. A-statistical summability of Fourier Series and Walsh–Fourier series. *Appl. Math. Inf. Sci.*, 2012. Vol. 6, no. 3. P. 535–538.
2. Aytar S. Rough statistical convergence. *Numer. Funct. Anal. Optim.*, 2008. Vol. 29, No. 3–4. P. 291–303. DOI: 10.1080/01630560802001064
3. Chawla M., Saroa M. S., Kumar V. On Λ-statistical convergence of order α in random 2-normed space. *Miskolc Math. Notes*, 2015. Vol. 16, No. 2. P. 1003–1015. DOI: 10.18514/MMN.2015.821
4. Çolak R., Bektaş Ç. A. λ-statistical convergence of order α. *Acta Math. Sci. Ser. B Engl. Ed.*, 2011. Vol. 31, No. 3. P. 953–959. DOI: 10.1016/S0252-9602(11)60288-9
5. Fast H. Sur la convergence statistique. *Colloq. Math.*, 1951. Vol. 2, No. 3–4. P. 241–244. URL: http://eudml.org/doc/209960
6. Fridy J. A. Statistical limit points. *Proc. Amer. Math. Soc.*, 1993. Vol. 118, No. 4. P. 1187–1192.
7. Gadjiev A. D., Orhan C. Some approximation theorems via statistical convergence. *Rocky Mountain J. Math.*, 2002. Vol. 32, No. 1. P. 129–138. DOI: 10.1216/rmjm/1030539612
8. Karakus S. Statistical convergence on probalistic normed spaces. *Math. Commun.*, 2007. Vol. 12, No. 1. P. 11–23.
9. Karakus S., Demirci K., Duman O. Statistical convergence on intuitionistic fuzzy normed spaces. *Chaos Solitons Fractals*, 2008. Vol. 35, No. 4. P. 763–769. DOI: 10.1016/j.chaos.2006.05.046
10. Maddox I. J. Statistical convergence in locally convex space. *Math. Proc. Cambridge Philos. Soc.*, 1988. Vol. 104, No. 1. P. 141–145. DOI: 10.1017/S0305004100065312
11. Maity M. *A Note on Rough Statistical Convergence.* 2016. 5 p. arXiv:1603.00180v1 [math.FA]
12. Maity M. *A Note on Rough Statistical Convergence of Order* α. 2016. 7 p. arXiv:1603.00183v1 [math.FA]
13. Malik P., Maity M. On rough convergence of double sequence in normed linear spaces. *Bull. Allahabad Math. Soc.*, 2013. Vol. 28, No. 1. P. 89–99.
14. Malik P., Maity M. On rough statistical convergence of double sequences in normed linear spaces. *Afr. Mat.*, 2016. Vol. 27, No. 1–2. P. 141–148. DOI: 10.1007/s13370-015-0332-9
15. Malik P., Maity M., Ghosh A. *Rough I-statistical Convergence of Sequences.* 2016. 21 p. arXiv:1601.03978v3 [math.FA]
16. Miller H. I. A measure theoretical subsequence characterization of statistical convergence. *Trans. Amer. Math. Soc.*, 1995. Vol. 347, No. 5. P. 1811–1819.
17. Mursaleen M., Noman A. K. On the spaces of λ-convergent and bounded sequences. *Thai J. Math.*, 2012. Vol. 8, No. 2. P. 311–329.
18. Pal S. K., Chandra D., Dutta S. Rough ideal convergence. *Hacet. J. Math. Stat.*, 2013. Vol. 42, No. 6. P. 633–640.
19. Phu H. X. Rough convergence in normed linear spaces. *Numer. Funct. Anal. Optim.*, 2001. Vol. 22, No. 1-2. P. 199–222. DOI: 10.1081/NFA-100103794
20. Phu H. X. Rough convergence in infinite dimensional normed spaces. *Numer. Func. Anal. Optim.*, 2003. Vol. 24, No. 3–4. P. 285–301. DOI: 10.1081/NFA-120022923

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 26–37
DOI: 10.15826/umj.2021.1.003

ON THE POTENTIALITY OF A CLASS OF OPERATORS RELATIVE TO LOCAL BILINEAR FORMS[1]

Svetlana A. Budochkina[†], **Ekaterina S. Dekhanova**[††]

Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya str., 117198 Moscow, Russia

[†]budochkina-sa@rudn.ru, [††]esdekhanova@gmail.com

Abstract: The inverse problem of the calculus of variations (IPCV) is solved for a second-order ordinary differential equation with the use of a local bilinear form. We apply methods of analytical dynamics, nonlinear functional analysis, and modern methods for solving the IPCV. In the paper, we obtain necessary and sufficient conditions for a given operator to be potential relative to a local bilinear form, construct the corresponding functional, i.e., found a solution to the IPCV, and define the structure of the considered equation with the potential operator. As a consequence, similar results are obtained when using a nonlocal bilinear form. Theoretical results are illustrated with some examples.

Keywords: Inverse problem of the calculus of variations, Local bilinear form, Potential operator, Conditions of potentiality.

1. Introduction

In the modern calculus framework, the classical inverse problem of the calculus of variations (IPCV) is a problem of constructing an integral functional such that its equations of extremals coincide with given equations. The issues considered in the paper are closely related to the following statement of the IPCV generalizing its classical statement. For a given equation, one needs to construct a functional such that its set of stationary points coincides with the set of solutions to this equation. These problems are also related to the mechanics of finite- and infinite-dimensional systems [7, 8, 11–13]. There is a large number of works devoted to IPCVs for different types of equations and their systems: in particular, for ordinary differential equations and differential equations with partial derivatives [4, 6, 13, 18, 19, 21], operator equations [2, 3, 14, 15], differential-difference equations [5, 9, 10], and stochastic differential equations [16, 17]. In these works, nonlocal bilinear forms were mainly used to solve an IPCV. Methods of investigating operators for the potentiality relative to local bilinear forms were developed in [6, 13, 20].

The main aim of the paper is to find a solution to an IPCV for a second-order ordinary differential equation. Local bilinear forms will play a significant role in the investigation.

Below, we use the notation and terminology of [2, 3, 13, 15].

Assume that U and V are linear normed spaces over $\mathbb{R}$.

The following definition and theorem will be needed for the sequel.

Definition 1 [13]. *An operator $N : D(N) \subset U \to V$ is called potential on the set $D(N)$ relative to a local bilinear form $\Phi(u;\cdot,\cdot) : V \times V \to \mathbb{R}$ if there exists a Gâteaux differentiable functional $F_N : D(F_N) = D(N) \to \mathbb{R}$ such that*

$$\delta F_N[u,h] = \Phi(u; N(u), h) \quad \forall u \in D(N), \quad \forall h \in D(N'_u). \tag{1.1}$$

[1]This paper was partially supported by the RUDN University Strategic Academic Leadership Program and by the Russian Foundation for Basic Research (project no. 19-08-00261a).

Theorem 1 [13]. *Consider a Gâteaux differentiable operator $N : D(N) \subset U \to V$ and a local bilinear form $\Phi(u;\cdot,\cdot) : V \times V \to \mathbb{R}$ such that, for any fixed elements $u \in D(N)$ and $g, h \in D(N'_u)$, the function $\psi(\varepsilon) = \Phi(u+\varepsilon h; N(u+\varepsilon h), g)$ belongs to the class $C^1[0,1]$. For N to be potential on the convex set $D(N)$ relative to Φ, it is necessary and sufficient to have*

$$\begin{gathered}\Phi\left(u; N'_u h, g\right) + \Phi'_u\left(h; N(u), g\right) = \Phi\left(u; N'_u g, h\right) + \Phi'_u\left(g; N(u), h\right)\\ \forall u \in D\left(N\right), \quad \forall h, g \in D\left(N'_u\right).\end{gathered} \tag{1.2}$$

Under this condition, the potential F_N is given as

$$F_N[u] = \int\limits_0^1 \Phi(u_0 + \lambda(u-u_0); N(u_0 + \lambda(u-u_0)), u - u_0)\, d\lambda + F_N[u_0], \tag{1.3}$$

where u_0 is a fixed element of $D(N)$.

Note that N'_u and Φ'_u are the Gâteaux derivatives of N and Φ at the point u.

2. Conditions of potentiality

Consider an ordinary differential equation of the second order

$$N(u) \equiv a(t,u(t))u''(t) + b(t,u(t))u'(t) + c(t,u(t))(u'(t))^2 + d(t,u(t)) = 0, \quad t \in [t_0, t_1]. \tag{2.1}$$

Here, $u = u(t)$ is an unknown function, $a \in C^2([t_0,t_1] \times T)$ and $b, c, d \in C^1([t_0,t_1] \times T)$ are given functions, and $T \subseteq \mathbb{R}$.

We define the domain of the operator N (2.1) as follows:

$$D(N) = \left\{u \in C^2[t_0,t_1] : u(t_0) = u_1,\ u(t_1) = u_2\right\}. \tag{2.2}$$

The domain $D(N'_u)$ consists of elements $h \in C^2[t_0,t_1]$ such that $(u+\varepsilon h) \in D(N)$ for all ε sufficiently small, i.e.,

$$D(N'_u) = \left\{h \in C^2[t_0,t_1] : h(t_0) = 0,\ h(t_1) = 0\right\}.$$

Let us introduce a local bilinear form

$$\Phi(u; v, g) = \int\limits_{t_0}^{t_1} M(t,u(t))v(t)g(t)\, dt, \tag{2.3}$$

where $M \in C^2([t_0,t_1] \times T)$, $M(t,u(t)) \neq 0$.

Theorem 2. *For the operator N (2.1) to be potential on $D(N)$ (2.2) relative to the local bilinear form (2.3), it is necessary and sufficient that the following conditions hold for all $u \in D(N)$ and all $t \in [t_0,t_1]$:*

$$a'_u(t,u(t))M(t,u(t)) + a(t,u(t))M'_u(t,u(t)) - 2c(t,u(t))M(t,u(t)) = 0, \tag{2.4}$$

$$a'_t(t,u(t))M(t,u(t)) + a(t,u(t))M'_t(t,u(t)) - b(t,u(t))M(t,u(t)) = 0. \tag{2.5}$$

P r o o f. We have

$$N'_u h = a'_u(t,u(t))u''(t)h(t) + a(t,u(t))h''(t) + b'_u(t,u(t))u'(t)h(t) + \\ + b(t,u(t))h'(t) + c'_u(t,u(t))(u'(t))^2 h(t) + 2c(t,u(t))u'(t)h'(t) + d'_u(t,u(t))h(t).$$

In this case, criterion (1.2) becomes

$$\int_{t_0}^{t_1} \Big(a'_u(t,u(t))M(t,u(t))u''(t)h(t)g(t) + a(t,u(t))M(t,u(t))h''(t)g(t) + \\ + b'_u(t,u(t))M(t,u(t))u'(t)h(t)g(t) + b(t,u(t))M(t,u(t))h'(t)g(t) + \\ + c'_u(t,u(t))M(t,u(t))(u'(t))^2 h(t)g(t) + 2c(t,u(t))M(t,u(t))u'(t)h'(t)g(t) + \\ + M(t,u(t))d'_u(t,u(t))h(t)g(t) + a(t,u(t))M'_u(t,u(t))u''(t)h(t)g(t) + \\ + b(t,u(t))M'_u(t,u(t))u'(t)h(t)g(t) + c(t,u(t))M'_u(t,u(t))(u'(t))^2 h(t)g(t) + \\ + M'_u(t,u(t))d(t,u(t))h(t)g(t) \Big) dt = \int_{t_0}^{t_1} \Big(a'_u(t,u(t))M(t,u(t))u''(t)h(t)g(t) + \\ + a(t,u(t))M(t,u(t))g''(t)h(t) + b'_u(t,u(t))M(t,u(t))u'(t)h(t)g(t) + \\ + b(t,u(t))M(t,u(t))g'(t)h(t) + c'_u(t,u(t))M(t,u(t))(u'(t))^2 h(t)g(t) + \\ + 2c(t,u(t))M(t,u(t))u'(t)g'(t)h(t) + M(t,u(t))d'_u(t,u(t))h(t)g(t) + \\ + a(t,u(t))M'_u(t,u(t))u''(t)h(t)g(t) + b(t,u(t))M'_u(t,u(t))u'(t)h(t)g(t) + \\ + c(t,u(t))M'_u(t,u(t))(u'(t))^2 h(t)g(t) + M'_u(t,u(t))d(t,u(t))h(t)g(t) \Big) dt \\ \forall u \in D(N), \quad \forall h, g \in D(N'_u),$$

or

$$\int_{t_0}^{t_1} \Big(a(t,u(t))M(t,u(t))h''(t)g(t) + b(t,u(t))M(t,u(t))h'(t)g(t) + \\ + 2c(t,u(t))M(t,u(t))u'(t)h'(t)g(t) \Big) dt = \int_{t_0}^{t_1} \Big(a(t,u(t))M(t,u(t))g''(t)h(t) + \\ + b(t,u(t))M(t,u(t))g'(t)h(t) + 2c(t,u(t))M(t,u(t))u'(t)g'(t)h(t) \Big) dt \\ \forall u \in D(N), \quad \forall h, g \in D(N'_u). \tag{2.6}$$

Integrating by parts and taking into consideration that $h, g \in D(N'_u)$, we obtain

$$\int_{t_0}^{t_1} \Big(a(t,u(t))M(t,u(t))h''(t)g(t) + b(t,u(t))M(t,u(t))h'(t)g(t) + \\ + 2c(t,u(t))M(t,u(t))u'(t)h'(t)g(t) \Big) dt = \int_{t_0}^{t_1} \Big(a''_{tt}(t,u(t))M(t,u(t))h(t)g(t) + \\ + 2a''_{tu}(t,u(t))u'(t)M(t,u(t))h(t)g(t) + 2a'_t(t,u(t))M'_t(t,u(t))h(t)g(t) + \\ + 2a'_t(t,u(t))M'_u(t,u(t))u'(t)h(t)g(t) + 2a'_t(t,u(t))M(t,u(t))h(t)g'(t) + \\ + a''_{uu}(t,u(t))(u'(t))^2 M(t,u(t))h(t)g(t) + a'_u(t,u(t))u''(t)M(t,u(t))h(t)g(t) + \\ + 2a'_u(t,u(t))u'(t)M'_t(t,u(t))h(t)g(t) + 2a'_u(t,u(t))M'_u(t,u(t))(u'(t))^2 h(t)g(t) +$$

$$
\begin{gathered}
+ 2a'_u(t,u(t))u'(t)M(t,u(t))h(t)g'(t) + 2a(t,u(t))M'_t(t,u(t))g'(t)h(t) + \\
+ 2a(t,u(t))M'_u(t,u(t))u'(t)h(t)g'(t) + a(t,u(t))M(t,u(t))h(t)g''(t) + \\
+ a(t,u(t))M''_{tt}(t,u(t))h(t)g(t) + 2a(t,u(t))M''_{tu}(t,u(t))u'(t)h(t)g(t) + \\
+ a(t,u(t))M''_{uu}(t,u(t))(u'(t))^2h(t)g(t) + a(t,u(t))M'_u(t,u(t))u''(t)h(t)g(t) - \\
- b'_t(t,u(t))M(t,u(t))h(t)g(t) - b'_u(t,u(t))u'(t)M(t,u(t))h(t)g(t) - \\
- b(t,u(t))M'_t(t,u(t))h(t)g(t) - b(t,u(t))M'_u(t,u(t))u'(t)h(t)g(t) - \\
- b(t,u(t))M(t,u(t))h(t)g'(t) - 2c'_t(t,u(t))M(t,u(t))u'(t)h(t)g(t) - \\
- 2c'_u(t,u(t))M(t,u(t))(u'(t))^2h(t)g(t) - 2c(t,u(t))(u'(t))^2M'_u(t,u(t))h(t)g(t) - \\
- 2c(t,u(t))M'_t(t,u(t))u'(t)h(t)g(t) - 2c(t,u(t))M(t,u(t))u''(t)h(t)g(t) - \\
- 2c(t,u(t))M(t,u(t))u'(t)h(t)g'(t)\Big)dt.
\end{gathered}
$$

Thus, equality (2.6) can be written in the form

$$
\begin{gathered}
\int_{t_0}^{t_1}\Big(a''_{tt}(t,u(t))M(t,u(t))h(t)g(t) + 2a''_{tu}(t,u(t))u'(t)M(t,u(t))h(t)g(t) + \\
+ 2a'_t(t,u(t))M'_t(t,u(t))h(t)g(t) + 2a'_t(t,u(t))M'_u(t,u(t))u'(t)h(t)g(t) + \\
+ 2a'_t(t,u(t))M(t,u(t))h(t)g'(t) + a''_{uu}(t,u(t))(u'(t))^2M(t,u(t))h(t)g(t) + \\
+ a'_u(t,u(t))u''(t)M(t,u(t))h(t)g(t) + 2a'_u(t,u(t))u'(t)M'_t(t,u(t))h(t)g(t) + \\
+ 2a'_u(t,u(t))M'_u(t,u(t))(u'(t))^2h(t)g(t) + 2a'_u(t,u(t))u'(t)M(t,u(t))h(t)g'(t) + \\
+ 2a(t,u(t))M'_t(t,u(t))g'(t)h(t) + 2a(t,u(t))M'_u(t,u(t))u'(t)h(t)g'(t) + \\
+ a(t,u(t))M''_{tt}(t,u(t))h(t)g(t) + 2a(t,u(t))M''_{tu}(t,u(t))u'(t)h(t)g(t) + \\
+ a(t,u(t))M''_{uu}(t,u(t))(u'(t))^2h(t)g(t) + a(t,u(t))M'_u(t,u(t))u''(t)h(t)g(t) - \\
- b'_t(t,u(t))M(t,u(t))h(t)g(t) - b'_u(t,u(t))u'(t)M(t,u(t))h(t)g(t) - \\
- b(t,u(t))M'_t(t,u(t))h(t)g(t) - b(t,u(t))M'_u(t,u(t))u'(t)h(t)g(t) - \\
- 2b(t,u(t))M(t,u(t))h(t)g'(t) - 2c'_t(t,u(t))M(t,u(t))u'(t)h(t)g(t) - \\
- 2c'_u(t,u(t))M(t,u(t))(u'(t))^2h(t)g(t) - 2c(t,u(t))(u'(t))^2M'_u(t,u(t))h(t)g(t) - \\
- 2c(t,u(t))M'_t(t,u(t))u'(t)h(t)g(t) - 2c(t,u(t))M(t,u(t))u''(t)h(t)g(t) - \\
- 4c(t,u(t))M(t,u(t))u'(t)h(t)g'(t)\Big)dt = 0 \\
\forall u \in D(N), \quad \forall h, g \in D\left(N'_u\right).
\end{gathered}
$$

Hence, we get

$$
\begin{gathered}
a''_{tt}(t,u(t))M(t,u(t)) + 2a'_t(t,u(t))M'_t(t,u(t)) + a(t,u(t))M''_{tt}(t,u(t)) - \\
- b'_t(t,u(t))M(t,u(t)) - b(t,u(t))M'_t(t,u(t)) = 0,
\end{gathered} \tag{2.7}
$$

$$
\begin{gathered}
2a''_{tu}(t,u(t))M(t,u(t)) + 2a'_t(t,u(t))M'_u(t,u(t)) + 2a'_u(t,u(t))M'_t(t,u(t)) + \\
+ 2a(t,u(t))M''_{tu}(t,u(t)) - 2c'_t(t,u(t))M(t,u(t)) - 2c(t,u(t))M'_t(t,u(t)) - \\
- b'_u(t,u(t))M(t,u(t)) - b(t,u(t))M'_u(t,u(t)) = 0,
\end{gathered} \tag{2.8}
$$

$$
a'_t(t,u(t))M(t,u(t)) + a(t,u(t))M'_t(t,u(t)) - b(t,u(t))M(t,u(t)) = 0, \tag{2.9}
$$

$$
\begin{gathered}
a(t,u(t))M''_{uu}(t,u(t)) + a''_{uu}(t,u(t))M(t,u(t)) + 2a'_u(t,u(t))M'_u(t,u(t)) - \\
- 2c'_u(t,u(t))M(t,u(t)) - 2c(t,u(t))M'_u(t,u(t)) = 0,
\end{gathered} \tag{2.10}
$$

$$
a'_u(t,u(t))M(t,u(t)) + a(t,u(t))M'_u(t,u(t)) - 2c(t,u(t))M(t,u(t)) = 0. \tag{2.11}
$$

Note that conditions (2.7)–(2.11) are reduced to (2.4) and (2.5). □

Remark 1. If $M = M(t)$, then

$$\Phi(v,g) = \int_{t_0}^{t_1} M(t)v(t)g(t)\,dt \tag{2.12}$$

is a nonlocal bilinear form and conditions (2.4) and (2.5) are represented in the form

$$a'_u(t,u(t)) - 2c(t,u(t)) = 0, \tag{2.13}$$
$$a'_t(t,u(t))M(t) + a(t,u(t))M'(t) - b(t,u(t))M(t) = 0. \tag{2.14}$$

Remark 2. If $M = M(t)$ and $a = a(t)$, $b = b(t)$, $c = c(t)$, then conditions (2.4) and (2.5) can be written in the form

$$c(t) = 0, \tag{2.15}$$
$$a'(t)M(t) + a(t)M'(t) - b(t)M(t) = 0. \tag{2.16}$$

Remark 3. If $M(t,u(t)) \equiv 1$, then

$$\Phi(v,g) = \int_{t_0}^{t_1} v(t)g(t)\,dt \tag{2.17}$$

and conditions (2.4) and (2.5) take the form

$$a'_u(t,u(t)) - 2c(t,u(t)) = 0, \tag{2.18}$$
$$a'_t(t,u(t)) - b(t,u(t)) = 0. \tag{2.19}$$

Remark 4. If $M(t,u(t)) \equiv 1$ and $a = a(t)$, $b = b(t)$, $c = c(t)$, then conditions (2.4) and (2.5) are reduced to

$$c(t) = 0, \tag{2.20}$$
$$a'(t) - b(t) = 0. \tag{2.21}$$

3. Finding a solution to the IPCV

Theorem 3. *If conditions (2.4) and (2.5) hold, then the corresponding functional is given as*

$$F_N[u] = \int_{t_0}^{t_1} \Big(-\frac{1}{2}M(t,u(t))a(t,u(t))(u'(t))^2 + B_M(t,u(t))\Big)\,dt, \tag{3.1}$$

where

$$B_M(t,u(t)) = \int_0^1 M(t,\tilde{u}(t,\lambda))d(t,\tilde{u}(t,\lambda))(u(t)-u_0(t))\,d\lambda + B_M(t,u_0(t)), \tag{3.2}$$

$\tilde{u}(t,\lambda) = u_0(t) + \lambda(u(t)-u_0(t))$, $u_0 = u_0(t)$ *is a fixed element of* $D(N)$, *and* $B_M \in C^2([t_0,t_1]\times T)$.

P r o o f. According to formula (1.3) and conditions (2.4) and (2.5) we have

$$
\begin{aligned}
&F_N[u]-F_N[u_0]=\\
&=\int_{t_0}^{t_1}\int_0^1\Big[M(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))\tilde u''_{tt}(t,\lambda)(u(t)-u_0(t))+\\
&+M(t,\tilde u(t,\lambda))b(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))+\\
&+M(t,\tilde u(t,\lambda))c(t,\tilde u(t,\lambda))(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))+\\
&+M(t,\tilde u(t,\lambda))d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\Big]\,d\lambda dt=\\
&=\int_{t_0}^{t_1}\int_0^1\Big[-M'_t(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))-\\
&-M'_{\tilde u(t,\lambda)}(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))-\\
&-M(t,\tilde u(t,\lambda))a'_t(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))-\\
&-M(t,\tilde u(t,\lambda))a'_{\tilde u(t,\lambda)}(t,\tilde u(t,\lambda))(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))-\\
&-M(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))'+\\
&+M(t,\tilde u(t,\lambda))b(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))+\\
&+M(t,\tilde u(t,\lambda))c(t,\tilde u(t,\lambda))(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))+\\
&+M(t,\tilde u(t,\lambda))d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\Big]\,d\lambda dt=\\
&=\int_{t_0}^{t_1}\int_0^1\Big[\tilde u'_t(t,\lambda)(u(t)-u_0(t))\Big(-M'_t(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))-\\
&-M(t,\tilde u(t,\lambda))a'_t(t,\tilde u(t,\lambda))+M(t,\tilde u(t,\lambda))b(t,\tilde u(t,\lambda))\Big)+\\
&+(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))\big(-M'_{\tilde u(t,\lambda)}(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))-\\
&-M(t,\tilde u(t,\lambda))a'_{\tilde u(t,\lambda)}(t,\tilde u(t,\lambda))+M(t,\tilde u(t,\lambda))c(t,\tilde u(t,\lambda))\big)-\\
&-M(t,\tilde u(t,\lambda))a(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))'+\\
&+M(t,\tilde u(t,\lambda))d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\Big]\,d\lambda dt=\\
&=\int_{t_0}^{t_1}\int_0^1\Big[-c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))-\\
&-a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))'+\\
&+M(t,\tilde u(t,\lambda))d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\Big]\,d\lambda dt.
\end{aligned}
\tag{3.3}
$$

Note that, using (2.4), we get

$$
\begin{aligned}
&\int_0^1\Big[-c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u'_t(t,\lambda))^2(u(t)-u_0(t))-\\
&-a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u'_t(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda=
\end{aligned}
$$

$$
\begin{gathered}
=\int_0^1 \Big[-c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))-\\
-a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)\frac{\partial \tilde u_t'(t,\lambda)}{\partial\lambda}\Big]\,d\lambda=\\
=\int_0^1 \Big[-c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))-\\
-\frac{\partial}{\partial\lambda}\left(a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2\right)+\\
+a'_{\tilde u(t,\lambda)}(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))+\\
+a(t,\tilde u(t,\lambda))M'_{\tilde u(t,\lambda)}(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))+\\
+a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda=\\
=\int_0^1 \Big[c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))-\\
-\frac{\partial}{\partial\lambda}\left(a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2\right)+\\
+a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda=\\
=\int_0^1 \Big[c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))+\\
+a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda-\\
-a(t,u(t))M(t,u(t))(u'(t))^2+a(t,u_0(t))M(t,u_0(t))(u_0'(t))^2.
\end{gathered}
$$

Hence,

$$
\begin{gathered}
\int_0^1 \Big[-c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))-\\
-a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda=\\
=\int_0^1 \Big[c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))+\\
+a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda-\\
-a(t,u(t))M(t,u(t))(u'(t))^2+a(t,u_0(t))M(t,u_0(t))(u_0'(t))^2
\end{gathered}
$$

and

$$
\begin{gathered}
\int_0^1 \Big[-c(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))(\tilde u_t'(t,\lambda))^2(u(t)-u_0(t))-\\
-a(t,\tilde u(t,\lambda))M(t,\tilde u(t,\lambda))\tilde u_t'(t,\lambda)(u(t)-u_0(t))'\Big]\,d\lambda=\\
=-\frac{1}{2}a(t,u(t))M(t,u(t))(u'(t))^2+\frac{1}{2}a(t,u_0(t))M(t,u_0(t))(u_0'(t))^2.
\end{gathered}
$$

Thus, (3.3) becomes

$$F_N[u]-F_N[u_0]=\int\limits_{t_0}^{t_1}\Big(-\frac{1}{2}a(t,u(t))M(t,u(t))(u'(t))^2+\frac{1}{2}a(t,u_0(t))M(t,u_0(t))(u_0'(t))^2+$$
$$+\int\limits_0^1 M(t,\tilde u(t,\lambda))d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\,d\lambda\Big)dt.$$

The use of (3.2) yields functional (3.1). □

Remark 5. If $M=M(t)$ and $a=a(t)$, $b=b(t)$, $c=c(t)$, then

$$F_N[u]=\int\limits_{t_0}^{t_1}\Big(-\frac{1}{2}M(t)a(t)(u'(t))^2+B_M(t,u(t))\Big)dt, \tag{3.4}$$

where

$$B_M(t,u(t))=\int\limits_0^1 M(t)d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\,d\lambda+B_M(t,u_0(t)). \tag{3.5}$$

Remark 6. If $M(t,u(t))\equiv 1$ and $a=a(t)$, $b=b(t)$, $c=c(t)$, then

$$F_N[u]=\int\limits_{t_0}^{t_1}\Big(-\frac{a(t)}{2}(u'(t))^2+B(t,u(t))\Big)dt,$$

where

$$B(t,u(t))=\int\limits_0^1 d(t,\tilde u(t,\lambda))(u(t)-u_0(t))\,d\lambda+B(t,u_0(t)).$$

4. The structure of variational equation (2.1)

Theorem 4. *Conditions* (2.4) *and* (2.5) *hold if and only if equation* (2.1) *takes the form*

$$\begin{aligned}N(u)\equiv a(t,u(t))u''(t)+\frac{1}{M(t,u(t))}\Big[M_t'(t,u(t))a(t,u(t))+M(t,u(t))a_t'(t,u(t))\Big]u'(t)+\\+\frac{1}{2M(t,u(t))}\Big[M_u'(t,u(t))a(t,u(t))+M(t,u(t))a_u'(t,u(t))\Big](u'(t))^2+\frac{(B_M)_u'(t,u(t))}{M(t,u(t))}=0.\end{aligned} \tag{4.1}$$

P r o o f. According to (1.1), for functional (3.1), we have

$$\delta F_N[u,h]=\int\limits_{t_0}^{t_1}\Big(-\frac{1}{2}M_u'(t,u(t))a(t,u(t))(u'(t))^2h(t)-$$
$$-\frac{1}{2}M(t,u(t))a_u'(t,u(t))(u'(t))^2h(t)-M(t,u(t))a(t,u(t))u'(t)h'(t)+(B_M)_u'(t,u(t))h(t)\Big)dt=$$
$$=\int\limits_{t_0}^{t_1}\Big(-\frac{1}{2}M_u'(t,u(t))a(t,u(t))(u'(t))^2h(t)-\frac{1}{2}M(t,u(t))a_u'(t,u(t))(u'(t))^2h(t)+$$

$$+ M'_t(t,u(t))a(t,u(t))u'(t)h(t) + M'_u(t,u(t))a(t,u(t))(u'(t))^2h(t) + M(t,u(t))a'_t(t,u(t))u'(t)h(t) +$$
$$+ M(t,u(t))a'_u(t,u(t))(u'(t))^2h(t) + M(t,u(t))a(t,u(t))u''(t)h(t) + (B_M)'_u(t,u(t))h(t)\Big)dt =$$
$$= \int\limits_{t_0}^{t_1}\Big(M(t,u(t))a(t,u(t))u''(t) + \big(M'_t(t,u(t))a(t,u(t)) + M(t,u(t))a'_t(t,u(t))\big)u'(t) +$$
$$+ \frac{1}{2}\big(M'_u(t,u(t))a(t,u(t)) + M(t,u(t))a'_u(t,u(t))\big)(u'(t))^2 + (B_M)'_u(t,u(t))\Big)h(t)dt = \Phi(u;N(u),h)$$
$$\forall u \in D(N), \quad \forall h \in D(N'_u).$$

Hence, equation (2.1) is represented in form (4.1).

On the other hand, equation (4.1) is derived from the stationarity condition of functional (3.1). This means that conditions (2.4) and (2.5) must be satisfied. □

5. Examples

Example 1. Consider the Emden–Fowler equation [1]

$$N(u) \equiv u''(t) + \frac{k_1}{t}u'(t) + k_2 t^{m-1}u^n(t) = 0, \quad t \in [t_0, t_1], \quad t_0 > 0, \tag{5.1}$$

where k_1, k_2, m, and n are constants, $n \in \mathbb{N}$.

In this case,

$$a = 1, \quad b(t) = \frac{k_1}{t}, \quad c = 0, \quad d(t,u(t)) = k_2 t^{m-1}u^n(t).$$

The operator N (5.1) is not potential on $D(N)$ (2.2) relative to bilinear form (2.17) because condition (2.21) is not satisfied.

We find $M = M(t)$ such that the operator N (5.1) is potential on $D(N)$ (2.2) relative to a bilinear form of type (2.12).

From condition (2.16), we obtain

$$M(t) = t^{k_1}.$$

Thus, the operator N (5.1) is potential on $D(N)$ (2.2) relative to the following bilinear form:

$$\Phi(v,g) = \int\limits_{t_0}^{t_1} t^{k_1}v(t)g(t)\,dt.$$

By formula (3.5), we get

$$B_M(t,u(t)) = \frac{k_2}{n+1}t^{m-1+k_1}u^{n+1}(t),$$

and functional (3.4) takes the form

$$F_N[u] = \int\limits_{t_0}^{t_1}\Big(-\frac{t^{k_1}}{2}(u'(t))^2 + \frac{k_2}{n+1}t^{m-1+k_1}u^{n+1}(t)\Big)\,dt. \tag{5.2}$$

Remark 7. The operator N of the Emden equation [8]

$$N(u) \equiv u''(t) + \frac{2}{t}u'(t) + u^5(t) = 0, \quad t \in [t_0, t_1], \quad t_0 > 0,$$

is potential on $D(N)$ (2.2) relative to the following bilinear form:

$$\Phi(v, g) = \int\limits_{t_0}^{t_1} t^2 v(t) g(t)\, dt$$

(see Example 1; $k_1 = 2$, $k_2 = 1$, $m = 1$, and $n = 5$).

In this case, functional (5.2) becomes

$$F_N[u] = \int\limits_{t_0}^{t_1} \Big(-\frac{t^2}{2}(u'(t))^2 + \frac{t^2}{6}u^6(t) \Big)\, dt. \tag{5.3}$$

Note that functional (5.3) was obtained in another way in [8].

Example 2. Consider the following equation:

$$N(u) \equiv 2tu''(t) + 2u'(t) + t(u'(t))^2 - u(t) - 1 = 0, \quad t \in [t_0, t_1]. \tag{5.4}$$

In this case,

$$a(t) = 2t, \quad b = 2, \quad c(t) = t, \quad d(u(t)) = -u(t) - 1.$$

The operator N (5.4) is not potential on $D(N)$ (2.2) relative to bilinear forms (2.12) and (2.17) because $c(t) \neq 0$.

We find $M = M(u(t))$ such that the operator N (5.4) is potential on $D(N)$ (2.2) relative to a bilinear form of type (2.3).

From conditions (2.4) and (2.5), we obtain

$$M(u(t)) = e^{u(t)}.$$

Thus, the operator N (5.4) is potential on $D(N)$ (2.2) relative to the following bilinear form:

$$\Phi(u; v, g) = \int\limits_{t_0}^{t_1} e^{u(t)} v(t) g(t)\, dt.$$

By formula (3.2), we get

$$B_M(u(t)) = -e^{u(t)}u(t),$$

and functional (3.1) takes the form

$$F_N[u] = \int\limits_{t_0}^{t_1} \Big(-te^{u(t)}(u'(t))^2 - e^{u(t)}u(t) \Big)\, dt.$$

Example 3. Consider the following equation [8]:

$$N(u) \equiv u''(t) - \frac{(u'(t))^2}{u(t)} + \frac{1}{u^2(t)} = 0, \quad t \in [t_0, t_1]. \tag{5.5}$$

Here,
$$a = 1, \quad b = 0, \quad c(u(t)) = -\frac{1}{u(t)}, \quad d(u(t)) = \frac{1}{u^2(t)}.$$

The operator N (5.5) is not potential on $D(N)$ (2.2) relative to bilinear forms (2.12) and (2.17) because conditions (2.13) and (2.18) do not hold.

We find $M = M(u(t))$ such that the operator N (5.5) is potential on $D(N)$ (2.2) relative to a bilinear form of type (2.3).

From conditions (2.4) and (2.5), we obtain
$$M(u(t)) = \frac{1}{u^2(t)}.$$

Thus, the operator N (5.5) is potential on $D(N)$ (2.2) relative to the following bilinear form:
$$\Phi(u; v, g) = \int_{t_0}^{t_1} \frac{1}{u^2(t)} v(t)g(t)\, dt.$$

By formula (3.2), we get
$$B_M(u(t)) = -\frac{1}{3u^3(t)},$$
and functional (3.1) takes the form
$$F_N[u] = \int_{t_0}^{t_1} \Big(-\frac{(u'(t))^2}{2u^2(t)} - \frac{1}{3u^3(t)}\Big)\, dt.$$

6. Conclusion

In the paper, we obtained the following results: the potentiality of the operator of a second-order ordinary differential equation relative to a local bilinear form was investigated, a formula for constructing the functional was given, and the structure of the corresponding Euler–Lagrange equation was defined. In particular, applications and extensions of the work consist in the possibility to establish connections between the invariance of the functional, the given equation, and its first integrals and to spread the proposed scheme of investigation to higher-order equations.

REFERENCES

1. Berkovich L. M. The generalized Emden-Fowler equation. *Symmetry in Nonlinear Mathematical Physics,* 1997. Vol. 1. P. 155–163.
2. Budochkina S. A., Savchin V. M. On direct variational formulations for second order evolutionary equations. *Eurasian Math. J.,* 2012. Vol. 3, No. 4. P. 23–34.
3. Budotchkina S. A., Savchin V. M. On indirect variational formulations for operator equations. *J. Funct. Spaces Appl.,* 2007. Vol. 5, No. 3. P. 231–242.
4. Filippov V. M. *Variatsionnye printsipy dlya nepotentsial'nykh operatorov* [Variational Principles for Non-potential Operators]. Moscow: PFU, 1985. 206 p. (in Russian)
5. Filippov V. M., Savchin V. M., Budochkina S. A. On the existence of variational principles for differential-difference evolution equations. *Proc. Steklov Inst. Math.,* 2013. Vol. 283. P. 20–34. DOI: 10.1134/S0081543813080038

6. Filippov V. M., Savchin V. M., Shorokhov S. G. Variational principles for nonpotential operators. *J. Math. Sci.,* 1994. Vol. 68, No. 3. P. 275–398. DOI: 10.1007/BF01252319
7. Galiullin A. S. *Obratnye zadachi dinamiki* [Inverse Problems of Dynamics]. Moscow: Nauka, 1981. 144 p. (in Russian)
8. Galiullin A. S., Gafarov G. G., Malayshka R. P., Khvan A. M. *Analiticheskaya dinamika sistem Gel'mgol'tsa, Birkgofa, Nambu* [Analytical dynamics of Helmholtz, Birkhoff, Nambu systems]. Moscow: Advances in Physical Sciences, 1997. 324 p. (in Russian)
9. Popov A. M. Potentiality conditions for differential-difference equations. *Differ. Equ.,* 1998. Vol. 34, No. 3. P. 423–426.
10. Popov A. M. Inverse problem of the calculus of variations for systems of differential-difference equations of second order. *Math. Notes,* 2002. Vol. 72, No. 5. P. 687–691. DOI: 10.1023/A:1021417324565
11. Santilli R. M. *Foundations of Theoretical Mechanics, I: The Inverse Problems in Newtonian Mechanics.* Berlin–Heidelberg: Springer–Verlag, 1977. 266 p.
12. Santilli R. M. *Foundations of Theoretical Mechanics, II: Birkhoffian Generalization of Hamiltonian Mechanics.* Berlin-Heidelberg: Springer–Verlag, 1983. 370 p. DOI: 10.1007/978-3-642-86760-6
13. Savchin V. M. *Matematicheskie metody mekhaniki beskonechnomernykh nepotentsial'nykh sistem* [Mathematical Methods of Mechanics of Infinite-Dimensional Nonpotential Systems]. Moscow: PFU, 1991. 237 p. (in Russian)
14. Savchin V. M. An operator approach to Birkhoff's equations. *Vestnik RUDN. Ser. Math.,* 1995. No. 2 (2). P. 111–123.
15. Savchin V. M., Budochkina S. A. On the structure of a variational equation of evolution type with the second t-derivative. *Differ. Equ.,* 2003. Vol. 39, No. 1. P. 127–134. DOI: 10.1023/A:1025184311701
16. Tleubergenov M. I., Ibraeva G. T. On the solvability of the main inverse problem for stochastic differential systems. *Ukrainian Math. J.,* 2019. Vol. 71, No. 1. P. 157–165. DOI: 10.1007/s11253-019-01631-w
17. Tleubergenov M. I., Ibraeva G. T. On inverse problem of closure of differential systems with degenerate diffusion. *Eurasian Math. J.,* 2019. Vol. 10, No. 2. P. 93–102.
18. Tonti E. On the variational formulation for linear initial value problems. *Ann. Mat. Pura Appl. (4),* 1973. Vol. 95. P. 331–359. DOI: 10.1007/BF02410725
19. Tonti E. Variational formulation for every nonlinear problem. *Internat. J. Engrg. Sci.,* 1984. Vol. 22, No. 11–12. P. 1343–1371. DOI: 10.1016/0020-7225(84)90026-0
20. Tonti E. Extended variational formulation. *Vestnik RUDN. Ser. Math.,* 1995. No. 2 (2). P. 148–162.
21. Tunitsky D. V. On the inverse variational problem for one class of quasilinear equations. *J. Geom. Phys.,* 2020. Vol. 148. P. 103568. DOI: 10.1016/j.geomphys.2019.103568

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 38–65

DOI: 10.15826/umj.2021.1.004

ON CHROMATIC UNIQUENESS OF SOME COMPLETE TRIPARTITE GRAPHS

Pavel A. Gein

Ural Federal University,
51, Lenin ave., Ekaterinburg, 620000, Russia
pavel.gein@gmail.com

Abstract: Let $P(G, x)$ be a chromatic polynomial of a graph G. Two graphs G and H are called chromatically equivalent iff $P(G, x) = H(G, x)$. A graph G is called chromatically unique if $G \simeq H$ for every H chromatically equivalent to G. In this paper, the chromatic uniqueness of complete tripartite graphs $K(n_1, n_2, n_3)$ is proved for $n_1 \geqslant n_2 \geqslant n_3 \geqslant 2$ and $n_1 - n_3 \leqslant 5$.

Keywords: Chromatic uniqueness, Chromatic equivalence, Complete multipartite graphs, Chromatic polynomial.

1. Introduction

All graphs in this paper are finite and simple, i.e., they do not contain loops and multiple edges. Basic terminology is used according to [1].

Let $G = (V, E)$ be a graph with a vertex set V and an edge set E. A *coloring* of the graph G with x colors is a map $\varphi\colon V \to \{1, 2, \ldots, x\}$ such that $\varphi(u) \neq \varphi(v)$ for any two adjacent vertices u and v of the graph G. We will call the numbers $1, 2, \ldots, x$ the colors. A graph is called *x-colorable* if there exists its coloring with x colors. The smallest integer x for which G is x-colorable is called the *chromatic number* of G and is denoted by $\chi(G)$. The number of colorings of a graph G with x colors is denoted by $P(G, x)$. It is well known (see, for example, [1]) that the function $P(G, x)$ is a polynomial in variable x. Two graphs G and H are called *chromatically equivalent* if $P(G, x) = P(H, x)$. A graph G is called *chromatically unique* if, for any graph H, the graphs G and H are chromatically equivalent iff they are isomorphic. Much attention of researches was drawn to the problem of chromatic uniqueness of complete multipartite graphs $K(n_1, n_2, \ldots, n_t)$. Here are some results, a more complete list can be found in the survey [16] and the monograph [6].

(1) A graph $K(n_1, n_2)$, where $n_1 \geqslant n_2 \geqslant 2$, is chromatically unique (see [11]).

(2) A graph $K(n_1, n_2, \ldots, n_{t-1}, 1)$ is chromatically unique iff $n_i \leqslant 2$ for all $i = 1, 2, \ldots t-1$ (see [14]).

(3) A graph $K(n_1, n_2, \ldots, n_t)$, where $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_t \geqslant 2$, is chromatically unique if $n_1 - n_t \leqslant 4$ (see [2, 4, 12, 13, 15]).

The main result of this paper is the following theorem.

Theorem 1. *A complete tripartite graph $K(n_1, n_2, n_3)$ is chromatically unique if $n_1 \geqslant n_2 \geqslant n_3 \geqslant 2$ and $n_1 - n_3 \leqslant 5$.*

The chromatic uniqueness of a graph $K(n_1, n_2, n_3)$ with $n_1 \geqslant n_2 \geqslant n_3 \geqslant 2$ and $n_1 - n_3 \leqslant 4$ was proved in [2, 12, 13]. The chromatic uniqueness of a graph $K(n_1, n_2, n_3)$ with $n_1 \geqslant n_2 \geqslant n_3 \geqslant 2$,

$n_1 - n_3 = 5$, and $n_1 + n_2 + n_3 \not\equiv 2 \pmod 3$ was proved in [9]. The main aim of this paper is to prove the theorem in the case $n_1 - n_3 = 5$ and $n_1 + n_2 + n_3 \equiv 2 \pmod 3$.

One of the most important tools for studying chromatic equivalence and chromatic uniqueness is chromatic invariants. Assume that a number is assigned to every graph. This number is called *a chromatic invariant* if it is the same for all chromatically equivalent graphs. For a chromatic invariant $\alpha(G)$ and two arbitrary graphs G_1 and G_2, let $\Delta\alpha(G_2, G_1) = \alpha(G_2) - \alpha(G_1)$. It is well known (see, for example, [1] or [7]) that the number of vertices, the number of edges, the number of connected components, and the number of triangles are chromatic invariants.

According to Zykov's theorem (see, for example, [1]), the chromatic polynomial can be written as

$$P(G, x) = \sum_{i=\chi}^{n} pt(G, i)x^{(i)},$$

where $pt(G, i)$ is the number of partitions of the vertex set of the graph G into t independent sets, and $x^{(i)}$ is the falling factorial of x, i.e., $x^{(i)} = x(x-1)\ldots(x-i+1)$. It follows from Zykov's theorem that the numbers $pt(G, i), i = \chi, \ldots, n$, are chromatic invariants. We are most interested in $pt(G, \chi + 1)$, which we will write simply as $pt(G)$.

The rest of the paper is structured as follows. Section 2 describes a connection between integer partitions and chromatic uniqueness of complete multipartite graphs and presents a schema of the proof. Also, some properties of chromatic invariants are discussed in this section. In Sections 3, 4, and 5, upper bounds of the invariant pt are proved. Section 6 contains the proof of the main theorem.

2. Integer partition lattice and chromatic invariants

Let n be a positive integer. A *partition of* n is a sequence of nonnegative integers $u = (u_1, u_2, \ldots,)$ such that $u_1 \geqslant u_2 \geqslant \ldots$ and $n = \sum_{i=1}^{\infty} u_i$. The *length* of a partition u is a number l such that $u_l > 0$ and $u_{l+1} = u_{l+2} = \ldots = 0$. Writing a partition, we will often omit its zero elements.

Let $u = (u_1, u_2, \ldots)$ and $v = (v_1, v_2, \ldots)$ be two partitions of a number n. Define $v \trianglelefteq u$ if

$$\begin{gathered} v_1 \le u_1, \\ v_1 + v_2 \le u_1 + u_2, \\ \ldots \\ v_1 + v_2 + \ldots + v_{t-1} \le u_1 + u_2 + \ldots + u_{t-1}, \end{gathered}$$

where t is the largest of the lengths of u and v. The relation $\trianglelefteq$ is called the *dominance order*. It was shown in [5] that all partitions of a number n form a lattice with respect to $\trianglelefteq$.

As was proved in [3], all partitions of a number n with fixed length form a lattice with respect to $\trianglelefteq$. Also, Baranskii and Sen'chonok introduced [3] a notion of *elementary transformation*. A partition $v = (v_1, v_2, \ldots v_t)$ is a result of an elementary transformation of a partition $u = (u_1, u_2, \ldots, u_t)$ if there exist indices i and j such that

(1) $1 \le i < j \le t$;

(2) $u_i - 1 \ge u_{i+1}$ and $u_{j-1} \ge u_j + 1$;

(3) $u_i - u_j = \delta \ge 2$;

(4) $v_i = u_i - 1, \quad v_j = u_j + 1, \quad u_k = v_k$ for all $k = 1, 2, \ldots, t, \quad k \ne i, j$.

It was proved in [3] that $v \trianglelefteq u$ holds if and only if the partition v can be obtained from the partition u by a with finite number of elementary transformations.

Every complete t-partite graph with n vertices can be identified with a partition of length t of the number n. Let $u = (u_1, u_2, \dots, u_t)$ be a partition of length t of the number n. We will write $K(u)$ instead of $K(u_1, u_2, \dots, u_t)$ and denote parts of the graph $K(u)$ by V_i, where $|V_i| = u_i$ for all $i = 1, 2, \dots, t$.

Let u be a partition of length t of a number n. We present the following schema for proving the chromatic uniqueness of the graph $K(u)$. By contradiction, we assume that the graph $K(u)$ is not chromatically unique. This means that there exists a graph H nonisomorphic to the graph $K(u)$, and the graphs H and $K(u)$ are chromatically equivalent. It is clear that the chromatic number of the graph H is t; thus, the graph H can be obtained from a complete t-partite graph $K(v)$ by removing a set of edges E. It was shown in [17] that different complete multipartite graphs are not chromatically equivalent; hence, E must be non-empty. Denote by E_{ij} the set of all edges $e \in E$ such that one end of e is in V_i and the other is in V_j and define $e_{ij} = |E_{ij}|$. For an arbitrary set of edges E from $K(v)$, denote by $\langle E \rangle$ the subgraph edge-induced by E.

The following lemma was proved in [8].

Lemma 1 [8, Lemma 1]. *Let $u = (u_1, \dots, u_i, \dots, u_j, \dots u_t) \to v = (\dots, u_i - 1, \dots u_j + 1, \dots)$ be an elementary transformation of a partition u, where $u_t \geq 2$. Then the graphs $K(u)$ and H are not chromatically equivalent.*

It is clear that every complete t-partite graph is t-colorable, but is not $(t-1)$-colorable; in other words, the chromatic number of a complete t-partite graph is equal to t. Let us compute $pt(K(u))$ for a complete multipartite graph $K(u_1, u_2, \dots, u_t)$. It is easy to show that any partition of the vertex set of the graph $K(u)$ into $t+1$ parts can be obtained by splitting exactly one part into two nonempty subsets; so $pt(K(u)) = \sum_{i=1}^{n} 2^{u_i - 1} - t$.

It was investigated in [4] how the invariant pt changes from the graph $K(v)$ to the graph H. Let us introduce all necessary definitions and auxiliary statements.

A complete multipartite subgraph G_1 of a graph $K(v)$ is called an *E-subgraph* if each part of the graph G_1 is contained in some part of the graph $K(v)$, and the edge set of the graph G_1 is contained in the set E. An arbitrary disjoint set of E-subgraphs is called *a garland*. We will say that a garland G' *destroys* a part V_i if every vertex of V_i is contained in some E-subgraph of the garland G'. A garland of cardinality p which destroys exactly $p-1$ parts is called *interesting*. The set of all edges of all E-subgraphs of a garland is called the *edge aggregate* and is denoted by $E(G)$. The set of all vertices of all E-subgraphs of a garland is called the *vertex aggregate* and is denoted by $V(G)$. A garland is called *k-edge* if its edge aggregate contains exactly k edges. The following properties were proved in [4].

(1) If the chromatic number of a graph H is equal to t and $pt(H) = 1$, then every garland of cardinality p destroys at most $p-1$ parts.

(2) Every garland is uniquely defined by its edge aggregate.

(3) $\Delta pt(H, K(v))$ is equal to the number of all interesting garlands.

The next lemma follows from these properties.

Lemma 2 [4, Corollary 2]. *If a graph H is obtained from a graph $K(v)$ by removing some set of edges E and the graphs $K(u)$ and H are chromatically equivalent, then $|E| \leq \Delta pt(H, K(v)) \leq 2^{|E|-1}$.*

An improvement of this estimate was obtained in [10]. A subgraph H' of a graph $\langle E \rangle$ is called a *coordinated subgraph of type* $K(s,1)$ if $H' \simeq K(s,1)$ and all s vertices of degree 1 are in the same part of $K(v)$. A part V_j of a graph $K(v)$ is called *active* if there exist a vertex $x \in V_j$ and an edge $e \in E$ such that x and e are incident.

Lemma 3 [10, Theorem]. *Let every active part of a graph $K(v)$ contain at least three vertices. If E induces a coordinated subgraph of type $K(|E|,1)$, then $\Delta pt(G,K(v)) = 2^{|E|} - 1$; otherwise, $\Delta pt(G,K(v)) \leqslant 2^{|E|-1} + 1$.*

Let $G' = \{G'_1, G'_2, \ldots, G'_p\}$ be a garland. We will say that the garland G' is of type $H_1 \dot\cup H_2 \dot\cup \ldots \dot\cup H_p$, where $\{H_1, H_2, \ldots H_p\}$ is a set of graphs, if $G'_i \simeq H_i$ for all $i = 1, 2, \ldots, p$. We will say that an edge set $F \subseteq E$ *induces a garland* (*an interesting garland*) if there exists a garland (an interesting garland, respectively) G such that $F = E(G)$. A set $F \subseteq E$ is called *continuable* if there exists a garland G such that $F \subseteq E(G)$. A set $F \subseteq E' \subseteq E$ is called *continuable outside* E' if there exists a garland G such that $E(G) \cap E' = F$. We will say that an edge e is in the garland G if $e \in E(G)$.

Let e be an arbitrary edge from E. Denote by $\xi_1(e)$ the number of triangles of the graph $K(v)$ containing the edge e. Let $\xi_1 = \sum_{e \in E} \xi_1(e)$.

Consider a triangle in a graph G that contains exactly two edges from E. Denote them by e_1 and e_2. We will call a subgraph induced by $\{e_1, e_2\}$ a Ξ_2*-subgraph*. Denote by ξ_2 the number of such subgraphs and by ξ_3 the number of triangles in $\langle E \rangle$.

Denote by $I_3(G)$ the number of triangles in the graph G. In [2], the following equality was established:

$$\Delta I_3(K(v), H) = \xi_1 - \xi_2 - 2\xi_3.$$

Note that, the removal of an edge cannot produce a new triangle; so $\Delta I_3(K(v), H)$ is equal to the number of triangles in $K(v)$ destroyed by removing an edge set E from $K(v)$.

3. Upper bound for the invariant pt in the case where $\langle E \rangle$ contains a triangle

The main goal of this section is, in the case where $\langle E \rangle$ contains a triangle, to obtain an upper bound for the number of interesting garlands better than the bound in Lemma 3. Denote by Δ the edge set of this triangle. One the way to achieve this is to count number of garlands whose edge aggregates contain an edge from Δ and an edge that is not in the triangle. The following lemma was proved in [10].

Lemma 4 [10, Lemma 3]. *If a set of edges $\{e_1, e_2, e_3\} = E_1 \subset E$ induces a triangle, then there is no nonempty set $\hat{E} \subseteq E \setminus E_1$ such that the sets $\{e_i\} \dot\cup E_1$ induce interesting garlands for all $i = 1, 2, 3$.*

We can deduce the following statement from this lemma.

Corollary 1. *If Δ induces a triangle, $\Delta \subseteq E$, and F is an arbitrary nonempty subset of $E \setminus \Delta$, then there exist at most three nonempty subsets $F' \subseteq \Delta$ such that $F \cup F'$ induces an interesting garland.*

P r o o f. Let us fix some $F \subseteq E \setminus \Delta$ and consider all possible nonempty subsets $F' \subseteq \Delta$. Define $\Delta = \{e_1, e_2, e_3\}$. Assume that $F \cup F'$ induces an interesting garland G'.

Assume that $|F'| = 2$ and, without loss of generality, let $F' = \{e_1, e_2\}$. Define $e_1 = xy$ and $e_2 = yz$. Note that the vertices x and z are in different parts of $K(v)$. Since $F \cup F'$ induces a garland and the edges e_1 and e_2 are adjacent, they are in the same E-subgraph of G'. Denote this E-subgraph by H_1. Since H_1 is a complete multipartite graph and the vertices x and z are in different parts, the edge xz is in the edge aggregate of G'. However, $xz \notin F'$ and $xz \in \Delta$; therefore, $xz \notin F$. Consequently, xz is not in the edge aggregate of G', a contradiction. Hence, $|F'| \neq 2$.

Thus, there are four choices for F': three one-element subsets and Δ. Using Lemma 4, one can deduce that at most two of the sets $F' \cup \{e_1\}$, $F' \cup \{e_2\}$, and $F' \cup \{e_3\}$ induce an interesting garland. Thus, there are at most three choices for F' and the corollary is proved. □

The following lemma describes all sets $F \subseteq E \setminus \Delta$ such that there exist exactly three subsets $F' \subseteq \{e_1, e_2, e_3\}$ such that $F \cup F'$ induces an interesting garland.

Lemma 5. *Assume that each part of $K(v)$ contains at least three vertices and edges $e_1, e_2, e_3 \in E$ induce a triangle. Let $F \subseteq E \setminus \{e_1, e_2, e_3\}$. If there are exactly three nonempty subsets $F' \subseteq \{e_1, e_2, e_3\}$ such that $F \cup F'$ induces an interesting garland, then F induces an interesting garland, $|F| > 1$, and there is no edge $f \in F$ such that f is adjacent to any of e_1, e_2, and e_3.*

P r o o f. Define $\Delta = \{e_1, e_2, e_3\}$. Fix some $F \subseteq E \setminus \Delta$ and assume that there exist three different subsets $F' \subseteq \Delta$ such that $F \cup F'$ induces an interesting garland. Denote this subsets by E_1, E_2, and E_3. Using Corollary 1, one can deduce that two of E_1, E_2, and E_3 are one-element sets and the third one is equal to Δ. Without loss of generality, let $E_1 = \{e_1\}, E_2 = \{e_2\}$, and $E_3 = \Delta$.

Let the set $E_1 \cup F$ induce an interesting garland G_1, the set $E_2 \cup F$ induce an interesting garland G_2, and the set $E_3 \cup F$ induce an interesting garland G_Δ. Denote the vertices of the triangle by x, y, and z.

Let us prove that an arbitrary edge $f \in F$ is not adjacent to any of the edges e_1, e_2, and e_3. By contradiction, assume that f is incident to some vertex from the triangle, without loss of generality, assume that f is incident to x. Denote the second end of f by w. Note that w and y or w and z are in different parts of $K(v)$. Without loss of generality, assume that w and y are in different parts of $K(v)$. Since the edges xy and $f = xw$ are adjacent and are in the same E-subgraph of the garland G_Δ, and w and y are in different parts of $K(v)$, the edge wy is in F. Therefore, since wx and wy are adjacent edges, they are in the same E-subgraph of the garland G_1. So, the edge xy is in the same E-subgraph. Thus, xy is in the edge aggregate of the garland G_1. By analogy, the edge xy is in the aggregate of the garland G_2. But this is impossible since $xy \in \Delta$ and the edge aggregates of the garlands G_1 and G_2 cannot contain a common edge from Δ.

Since any edge from F is not adjacent to any edge from Δ, the set $\{e_1\}$ induces a E-subgraph of the garland G_1. Therefore, the graph $\langle F \rangle$ consists of E-subgraphs, so F induces a garland. Denote this garland by G_F.

Note that all the garlands G_1, G_2 and G_Δ have the same cardinality. Denote this cardinality by p. Since all this garlands are interesting, each of them destroys exactly $p-1$ parts of $K(v)$. Note also that $|G_F| = p - 1$.

Let $x \in V_i$, $y \in V_j$, $z \in V_k$, $e_1 = xy$, and $e_2 = yz$. Note that $V(G_\Delta) = V(G_1) \cup V(G_2)$. This means that if G_1 or G_2 destroy some part, then G_Δ also destroys this part. Note also that if one of the garlands G_1, G_2, and G_Δ destroys some part V other than V_i, V_j, and V_k, then G_F also destroys V. Since no edge from $E(G_1)$ is incident to z, the garland G_1 cannot destroy V_k. If G_2 destroys V_k, then G_Δ also destroys V_k and thus G_Δ destroys more parts than G_1 does, a contradiction. Therefore, G_2 does not destroy V_k. If G_2 does not destroy V_j, then G_F destroys $p-1$ parts and $|G_F| = p-1$, a contradiction. So G_2 destroys V_j. Since no edge from F is incident to y, the garland G_F cannot destroy V_j, thus G_F destroys one part less than G_2 does. Therefore, G_F destroys $p-1$ parts, and G_F is an interesting garland.

It remains to prove that $|F| > 1$. If $|F| = 1$, then G_Δ is a garland of type $K(1,1,1)\dot{\cup}K(1,1)$. Note that such a garland cannot destroy any part, since each active part contains at least three vertices. □

Now we are ready to prove a better upper bound in the case where $\langle E \rangle$ contains a triangle.

Lemma 6. *Assume that each part of $K(v)$ contains at least three vertices. If $\langle E\rangle$ contains a triangle, then the number of interesting garland does not exceed* $2^{|E|-2} + 2^{|E|-3} - 3|E| + 13$.

P r o o f. Denote edges of the triangle by e_1, e_2, and e_3, and let $\Delta = \{e_1, e_2, e_3\}$. Define $E' = E \setminus \Delta$.

Case 1. Assume that E' induces a coordinated subgraph of type $K(|E|-3, 1)$. Then all edges of E are located in one of the ways shown in Fig. 1.

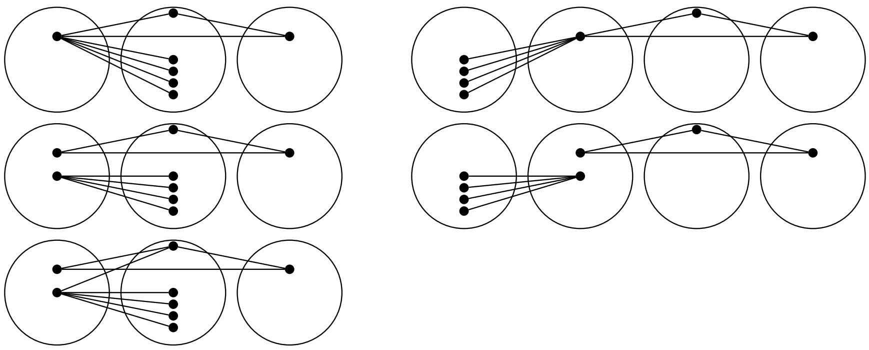


Figure 1. Triangle and a coordinated subgraph of type $K(|E|-3, 1)$.

In the first case, there are $2 + 1 + 2^{|E|-2} - 1 = 2^{|E|-2} + 2$ interesting garlands.
In the second case, there are $4 + 2^{|E|-3} - 1 = 2^{|E|-3} + 3$ interesting garlands.
In the third case, there are at most $4 + 2^{|E|-3} - 1 + 3 = 2^{|E|-3} + 6$ interesting garlands.
In the fourth case, there are $4 + 2^{|E|-3} - 1 = 2^{|E|-3} + 3$ interesting garlands.
In the fifth case, there are at most $4 + 2^{|E|-3} - 1 + 1 = 2^{|E|-3} + 4$ interesting garlands.
In the sixth case, there are at most $2^{|E|-3} + 6 < 2^{|E|-2} + 2^{|E|-3} - 3|E| + 13$ interesting garlands.
In all the cases, the number of interesting garlands does not exceed $2^{|E|-3} + 6 < 2^{|E|-2} + 2^{|E|-3} - 3|E| + 13$.

Case 2. Now consider the case where E' does not induce a coordinated subgraph of type $K(|E|-3, 1)$.

There are four interesting garlands G such that $E(G) \subseteq \Delta$. By Lemma 3, there exist at most $2^{|E|-4} + 1$ interesting garlands G such that $E(G) \subseteq E'$.

Estimate the number of interesting garlands whose edge aggregates contain some edge from E' and some edge from Δ. Define $F = E(G), F_\Delta = F \cap \Delta$, and $F' = F \cap E'$. By Lemma 5, if for F there exist three F' such that $F \cup F'$ induces an interesting garland, then F' induce an interesting garland, $|F'| > 1$, and any edge from F' is not adjacent to any edge from Δ. Denote the number of such F by X.

Case 2.1. Assume that any edge from E' is not adjacent to any edge from Δ. Using Lemma 3 and the fact that $|F| > 1$, one can deduce that

$$X \leqslant 2^{|E|-4} + 1 - (|E| - 3) = 2^{|E|-4} - |E| + 4.$$

For any other $F' \subseteq E', |F'| > 1$, there are at most two subsets $F_\Delta \subseteq \Delta$ such that $F' \cup F_\Delta$ induces an interesting garland. If $F' \subseteq E'$ and $|F'| = 1$, there exist no $F_\Delta \subseteq \Delta$ such that $F' \cup F_\Delta$ induces an interesting garland. Therefore, the number of interesting garlands does not exceed

$$\begin{aligned} 3X + 2(2^{|E|-3} - 1 - X - (|E| - 3)) + 2^{|E|-4} + 1 + 4 &= \\ = 3X + 2^{|E|-2} - 2 - 2X - 2|E| + 6 + 2^{|E|-4} + 1 + 4 &= \\ = 2^{|E|-2} + X - 2|E| + 2^{|E|-4} + 9 &\leqslant \\ \leqslant 2^{|E|-2} + 2^{|E|-4} - |E| + 4 - 2|E| + 2^{|E|-4} + 9 &= 2^{|E|-2} + 2^{|E|-3} - 3|E| + 13. \end{aligned}$$

Case 2.2. Now consider the case when there are exactly $k \geqslant 1$ edges from E' such that each of them is adjacent to some edge from Δ. Denote the set of this edges by H.

If for a set $F' \subseteq E'$ there exists exactly three sets $F_\Delta \subseteq \Delta$ such that $F' \cup F_\Delta$ induces an interesting garland, then $F' \subseteq E' \setminus H$ and $|F'| > 1$, so

$$X \leqslant 2^{|E|-3-k} - (|E| - 3 - k) - 1.$$

For any other $F' \subseteq E', |F'| > 1$, there are at most two such sets F_Δ. If $|F'| = 1$, then if $F' \subseteq H$, then there is at most one such F_Δ and if $F' \not\subseteq H$, then such F_Δ does not exist. Therefore, the number of interesting garland does not exceed

$$\begin{aligned} 3X + 2(2^{|E|-3} - 1 - X - (|E| - 3)) + k + 2^{|E|-4} + 1 + 4 &= \\ 3X + 2^{|E|-2} - 2 - 2X - 2|E| + 6 + 2^{|E|-4} + 1 + 4 + k &= \\ = 2^{|E|-2} + X - 2|E| + 2^{|E|-4} + 9 + k &\leqslant \\ \leqslant 2^{|E|-2} - 2|E| + 2^{|E|-4} + 9 + 2^{|E|-3-k} - (|E| - 3 - k) - 1 + k &= \\ = 2^{|E|-2} + 2^{|E|-4} - 3|E| + 11 + 2^{|E|-3-k} + 2k. & \end{aligned}$$

It remains to prove that

$$2^{|E|-2} + 2^{|E|-4} - 3|E| + 11 + 2^{|E|-3-k} + 2k \leqslant 2^{|E|-2} + 2^{|E|-3} - 3|E| + 13.$$

It is sufficient to prove that

$$2^{|E|-3-k} + 2k \leqslant 2^{|E|-3} + 2.$$

Let $m = |E| - 3$. Consider the function $f(x) = 2^{m-x} + 2x$ defined on the closed interval from 1 to m. Calculate the first and second derivatives:

$$\begin{aligned} f'(x) &= -2^{m-x} \ln 2 + 2; \\ f''(x) &= 2^{m-x} \ln^2 2 > 0. \end{aligned}$$

Therefore, f is a convex function, so it takes its maximum value at the end point of the interval. So it suffices to verify the inequality for $k = 1$ and $k = m = |E| - 3$. For $k = 1$, the inequality takes the form

$$2^{|E|-3} + 2 \leqslant 2^{|E|-3} + 2$$

and is valid. Consider the case $k = m = |E| - 3$. In this case, we need to prove that $2m \leqslant 2^m + 2$. Dividing both sides by 2, one can obtain $m - 1 \leqslant 2^{m-1}$. By Bernulli's inequality,

$$2^{m-1} = (1 + 1)^{m-1} \geqslant 1 + m - 1 = m > m - 1,$$

and the lemma is proved. □

In some cases, we can use a better upper bound than Lemma 3 gives. In such cases, the following lemma is useful.

Lemma 7. *Assume that each part of $K(v)$ contains at least three vertices and edges $e_1, e_2, e_3 \in E$ induce a triangle. Let $E' = E \setminus \{e_1, e_2, e_3\}$. If there are X interesting garlands G such that $E(G) \subseteq E'$, Y nonempty subsets of E' continuable outside E', k edges $h \in E'$ for each of which there exists an edge $h' \in \Delta$ such that h and h' induce a coordinated subgraph of type $K(2,1)$, then the number of interesting garlands doest not exceed $2X + 2Y - 5|E| + 19 + k$.*

P r o o f. Consider an arbitrary subset $F \subseteq E'$ and estimate the number of interesting garlands G such that $F \subseteq E(G)$ and $E(F) \cap \{e_1, e_2, e_3\} \neq \varnothing$. Let $H = E(F) \cap \{e_1, e_2, e_3\}$. Note that F is a nonempty subset of E' continuable outside E'.

Case 1. $|F| = 1$. Let $F = \{f\}$. If f does not induce a coordinated subgraph of type $K(2,1)$ with any of edges from $\{e_1, e_2, e_3\}$, then H must be an E-subgraph. This means that either H is a one-element set or $H = \{e_1, e_2, e_3\}$. In both cases, $F \cup H$ does not induce an interesting garland. Therefore, f must induce a coordinated subgraph of type $K(2,1)$ with some edge from $\{e_1, e_2, e_3\}$. By the lemma hypothesis, there are k such edges. Note that, in this case, H must be a one-element set and it is uniquely defined by the edge f. So, in this case, there are at most k interesting garlands.

Case 2. $|F| > 1$ and F does not induce interesting garlands. By Lemma 5, there are at most two subsets H. Note that F can be chosen in

$$Y - X - (|E| - 3) = Y - X - |E| + 3$$

ways; so, in this case, the number of interesting garlands does not exceed

$$2(Y - X - |E| + 3) = 2Y - 2X - 2|E| + 6.$$

Case 3. $|F| > 1$ and F induces interesting garlands. In this case, by Lemma 7, there are at most three sets H, so the number of interesting garlands does not exceed

$$3(X - (|E| - 3)) = 3X - 3|E| + 9.$$

There are X interesting garlands whose edge aggregates are in the set E' and four interesting garlands whose edge aggregates are in the set $\{e_1, e_2, e_3\}$. Therefore, the number of interesting garlands does not exceed

$$2Y - 2X - 2|E| + 6 + 3X - 3|E| + 9 + k + X + 4 = 2Y + 2X - 5|E| + 19,$$

and the lemma is proved. □

Lemma 8. *Assume that each active part of $K(v)$ contains at least three vertices. If $\langle E \rangle$ contains two triangles that do not have a common edge, then the number of interesting garlands does not exceed $2^{|E|-2} - 12|E| + 58$.*

P r o o f. Denote by Δ the edge set of one triangle, and let $E' = E \setminus \Delta$. Let X be the number of interesting garlands G such that $E(G) \subseteq E'$, and let Y be the number of nonempty subsets of E' continuable outside E'.

By Lemma 6,

$$X \leqslant 2^{|E|-5} + 2^{|E|-6} - 3(|E| - 3) + 13 = 3 \cdot 2^{|E|-6} - 3|E| + 22.$$

Denote by Δ' the edge set of the other triangle. To estimate the number of nonempty subsets of E' continuable outside E', consider an arbitrary $F \subseteq E' \setminus \Delta'$ and estimate the number of sets

$F' \subseteq \Delta'$ such that $F \cup F'$ is a continuable set. Note that if $F \cup F'$ is a continuable set, then F' is either empty, or a one-element set, or Δ'. Since F can be chosen in $2^{|E|-6}$ ways, there are at most $5 \cdot 2^{|E|-6}$ continuable sets; thus, there are at most $5 \cdot 2^{|E|-6}$ nonempty continuable sets.

Using Lemma 7 and assuming $k \leqslant |E| - 3$, one can conclude that the number of interesting garlands does not exceed

$$2 \cdot (3 \cdot 2^{|E|-6} - 3|E| + 22) + 2 \cdot (5 \cdot 2^{|E|-6} - 1) - 5|E| + 19 + |E| - 3 =$$
$$= 16 \cdot 2^{|E|-6} - 12|E| + 55 = 2^{|E|-2} - 12|E| + 58.$$

□

4. Upper bound for the invariant pt in the case where $\langle E \rangle$ contains a Ξ_2-subgraph

The main goal of this section is, in the case where $\langle E \rangle$ contains a Ξ_2-subgraph, to obtain a better upper bound for the number of interesting garlands. Recall that a pair of edges $e, f \in E$ induces a Ξ_2-subgraph if it induces a triangle in $K(v)$ whose third edge is not in E.

Lemma 9. *If edges f and g induce a Ξ_2-subgraph, then there is no a garland G such that $f, g \in E(G)$.*

P r o o f. Define $f = xy$ and $g = yz$ and note that the vertices x and z are in different parts of $K(v)$.

By contradiction, assume that there is a garland G such that $f, g \in E(G)$. The edges f and g are adjacent, so they are in the same E-subgraph H' of the garland G. Since the vertices y and z are in different parts of $K(v)$, the edge yz must be in the subgraph H', so $yz \in E$, a contradiction. □

Now we are ready to prove a better upper bound.

Lemma 10. *Let $|E| \geqslant 4$ and each active part of $K(v)$ contains at least three vertices. If $\langle E \rangle$ contains a Ξ_2-subgraph, then the number of interesting garlands does not exceed $2^{|E|-1}$.*

P r o o f. Let edges f and g induce a Ξ_2-subgraph. Consider the cases.

Case 1. Assume that $\langle E \rangle$ contains a coordinated subgraph of type $K(|E|-1, 1)$. Without loss of generality, assume that f is not in the coordinated subgraph of type $K(|E|-1, 1)$. Then there are two possible configurations for E, they are shown in Fig. 2.

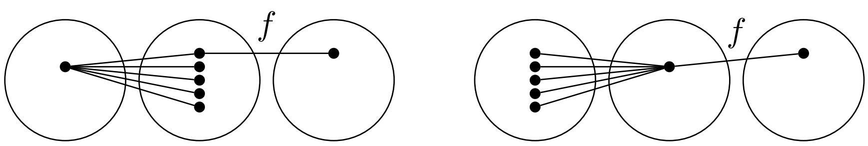


Figure 2. Ξ_2-subgraph and coordinated subgraph of type $K(|E|-1, 1)$.

In both cases, the edge f is in exactly one interesting garland. There are exactly $2^{|E|-1} - 1$ interesting garlands that do not contain the edge f. Thus, there are exactly

$$2^{|E|-1} - 1 + 1 = 2^{|E|-1}$$

interesting garlands.

Case 2. Now consider the case where $\langle E\rangle$ does not contain a coordinated subgraph of type $K(|E|-1,1)$. By Lemma 9, the edge aggregate of any garland is in the set $E\setminus\{f\}$ or in the set $E\setminus\{e\}$. By Lemma 3, there are at most $2^{|E|-2}+1$ interesting garlands whose edge aggregates are in the set $E\setminus\{f\}$ and at most $2^{|E|-2}+1$ interesting garlands whose edge aggregates are in the set $E\setminus\{e\}$. Note that $|E|-2$ interesting one-edge garlands were counted twice, so the number of interesting garlands does not exceed

$$2\cdot(2^{|E|-2+1})-(|E|-2)=2^{|E|-1}-|E|+4\leqslant 2^{|E|-1},$$

and the lemma is proved. □

In the case where $\langle E\rangle$ does not contain a coordinated subgraph of type $K(|E|-1,1)$, we can obtain a better upper bound.

Lemma 11. *Let $|E|\geq 5$, and let each part of $K(v)$ contain at least three vertices. If $\langle E\rangle$ does not contain a coordinated subgraph of type $K(|E|-1,1)$ and contains a Ξ_2-subgraph, then the number of interesting garlands does not exceed $2^{|E|-1}-|E|+1$.*

P r o o f. By Lemma 6, it is sufficient to consider the case where $\langle E\rangle$ does not contain a triangle.

Let edges e and g induce a Ξ_2-subgraph. Note that, by Lemma 9, the edge aggregate of any garland cannot contain both the edges e and g.

Case 1. Assume that the edge g is in two different Ξ_2-subgraphs. Define $E_e=E\setminus\{g\}$ and note that if $e\in E(G)$, then $E(G)\subseteq E_e$. By the lemma hypothesis, E_e does not induce a coordinated subgraph of type $K(|E|-1,1)$, so, by Lemma 3, there are at most $2^{|E|-2}+1$ interesting garlands whose edge aggregates are subsets of E_e. In this case, it remains to estimate the number of interesting garlands whose edge aggregates contain the edge g.

Let $f\neq e$ be an edge such that the edges f and g induce a Ξ_2-subgraph. Define $E_g=E\setminus\{f,e\}$ and note that if the edge g is in G, then $E(G)\subseteq E_g$.

Case 1.1. Assume that E_g induces a coordinated subgraph of type $K(|E_g|,1)$. Then there are exactly $2^{|E|-2}-1$ interesting garlands whose edge aggregates are subsets of E_g. Recall that there are at most $2^{|E|-2}+1$ interesting garlands whose edge aggregates are subsets of E_e. Note also that there are $2^{|E|-3}$ interesting garlands whose edge aggregates are subsets of $E\setminus\{e,g\}$ (which are $2^{|E|-3}-1$ interesting garlands of type $K(s,1)$ whose edge aggregates are subsets of $E\setminus\{e,g,f\}$ and a one-edge garland induced by f). All these garlands were counted twice, so the number of interesting garlands does not exceed

$$2^{|E|-2}-1+2^{|E|-2}+1-2^{|E|-3}=2^{|E|-1}-2^{|E|-3}.$$

It remains to prove that

$$2^{|E|-1}-2^{|E|-3}\leqslant 2^{|E|-1}-|E|+1.$$

To do this, it suffices to prove that the inequality $n-1\leqslant 2^{n-3}$ holds for all integer $n\geqslant 5$. Let us prove this by induction on n.

The base case. If $n=5$, then $5-1=4=2^2$ and the inequality is verified.

The induction step. Assume that the inequality is proved for n and prove it for $n+1$. We need to prove $2^{n-2}\geqslant n$. By the induction hypothesis,

$$2^{n-2}=2\cdot 2^{n-3}\geqslant 2\cdot(n-1)=n+n-2\geqslant n+3>n,$$

and the inequality is proved.

Case 1.2. Now consider the case where E_g does not induce a coordinated subgraph of type $K(|E_g|,1)$. In this case, by Lemma 3, there are at most $2^{|E|-3}+1$ interesting garlands whose edge aggregates are subsets of E_g. Note also that there are $|E|-3$ interesting one-edge garlands induced by edges from $E\setminus\{e,g\}$. So, the number of interesting garlands does not exceed

$$2^{|E|-3}+1+2^{|E|-2}+1-(|E|-3)=3\cdot 2^{|E|-3}+5.$$

It remains to prove that

$$3\cdot 2^{|E|-3}-|E|+5\leqslant 2^{|E|-1}-|E|+1.$$

It suffices to check that

$$3\cdot 2^{|E|-3}+4\leqslant 2^{|E|-1}=4\cdot 2^{|E|-3},$$

which is equivalent to $4\leqslant 2^{|E|-3}$, which is true because $|E|\geqslant 5$.

Case 2. Now consider the case where the edges f and g are in only one Ξ_2-subgraph. Define $E_g=E\setminus\{e\}$ and $E_e=E\setminus\{e\}$. Note that, by Lemma 3, there are at most $2^{|E|-1}+1$ interesting garlands whose edge aggregates are subsets of E_g and at most $2^{|E|-1}+1$ interesting garlands whose edge aggregates are subsets of E_e.

Assume that there exist at least $|E|+1$ interesting garlands whose edge aggregates are subsets of $E\setminus\{e,g\}$. In this case, there are at most

$$2(2^{|E|-1}+1)-(|E|+1)=2^{|E|}+2-|E|-1=2^{|E|}-|E|+1$$

interesting garlands. Therefore, it suffices to prove the lemma in the case when there are at most $|E|$ garlands whose edge aggregates are subsets of $E\setminus\{e,g\}$.

Note that, in any case, there are $|E|-2$ interesting one-edge garlands induced by edges from $E\setminus\{e,g\}$. Consequently, there are at most two interesting garlands whose edge aggregates are subsets of $E\setminus\{e,g\}$ and contain at least two edges.

Note that it suffices to prove that the edge f is in at most $2^{|E|-2}-|E|$ interesting garlands. Indeed, in this case, since there are at most $2^{|E|-1}+1$ interesting garland whose edge aggregates are subsets of E_g, the total number of interesting garlands does not exceed

$$2^{|E|-2}-|E|+2^{|E|-2}+1=2^{|E|-1}-|E|+1.$$

By analogy, it suffices to prove that the edge g is in at most $2^{|E|-2}-|E|$ interesting garlands.

To prove this, we build either $|E|$ subsets $F\subseteq E\setminus\{e,f\}$ such that $F\cup\{f\}$ does not induce an interesting garland or $|E|$ subsets $F\subseteq E\setminus\{e,f\}$ such that $F\cup\{g\}$ does not induce an interesting garland.

Case 2.1. Assume that the edge e is in three different coordinated subgraphs of type $K(2,1)$. In this case, either $E\setminus\{e,g\}$ contains a coordinated subgraph of type $K(3,1)$ and so contains at least three interesting garlands with more than one edge or g is in two different Ξ_2-subgraphs. Both cases are impossible.

Case 2.2. Assume that the edge e is in two different coordinated subgraphs of type $K(2,1)$. Let edges h_1 and h_2 be such that the pairs of edges $\{e,h_1\}$ and $\{e,h_2\}$ induce Ξ_2-subgraphs. Note that, in this case, the edge e is in the coordinated subgraph of type $K(3,1)$ because, otherwise, the edge g is in two Ξ_2-subgraphs.

Consider an arbitrary edge $\hat{h}\in E\setminus\{e,g,h_1,h_2\}$. Note that the set $\{e,\hat{h}\}$ does not induce a coordinated subgraph of type $K(2,1)$; the set $\{e,h_1,\hat{h}\}$ does not induce an interesting garland, since an interesting three-edge garland must be of type $K(3,1)$, a triangle, or $K(2,1)\dot{\cup}K(1,1)$, and all cases are impossible. By analogy, the set $\{e, h_2, \hat{h}\}$ does not induce an interesting garland. So, we have built

$$3 \cdot (|E| - 4) = 3|E| - 12 = |E| + 2(|E| - 6)$$

necessary subsets. If $|E| \geqslant 6$, then the lemma is proved.

Consider the case $|E| = 5$. Denote by g' a single edge from $E \setminus \{e, g, h_1, h_2\}$. Note, that the sets $\{g, h_1\}, \{g, h_2\}, \{g, h_1, h_2\}, \{g, g', h_1\}$, and $\{g, g', h_2\}$ do not induce interesting garlands (see Fig. 3) and, in this case, the lemma is proved.

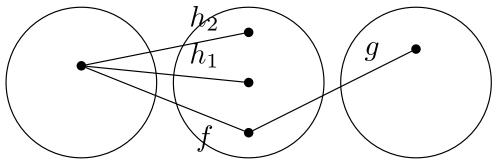


Figure 3. Cordinated subgraph of type $K(3, 1)$ and Ξ_2-subgraph.

Case 2.3. Assume that the edge e is in exactly one coordinated subgraph of type $K(2, 1)$. Denote the second edge of this subgraph by h.

Case 2.3.1. Assume that the edge g is not in any coordinated subgraph of type $K(2, 1)$. Consider an arbitrary edge $f \in E \setminus \{e, g\}$. Note that the pair of edges $\{g, f\}$ does not induce a coordinated subgraph of type $K(2, 1)$. Note also that, for an arbitrary edge $f \in E \setminus \{e, g, h\}$, the triple of edges $\{g, h, f\}$ does not induce an interesting garland. Therefore, we have built

$$|E| - 2 + |E| - 3 = 2|E| - 5 \geqslant |E|$$

necessary subsets for the edge g.

Case 2.3.2. Assume that the edge g is in the coordinated subgraph of type $K(2, 1)$. By the previous cases, it suffices to consider the case where g is in exactly one coordinated subgraph of type $K(2, 1)$. Denote the second edge of this subgraph by h'. Note that the edges e, g, h, and h' form one of the two configurations shown in Fig. 4.

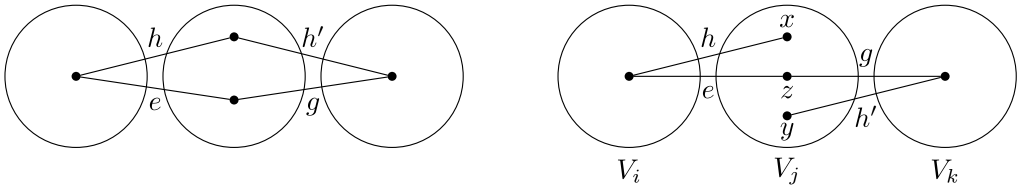


Figure 4. Ξ_2-subgraph and two coordinated subgraphs of type $K(2, 1)$.

Consider an arbitrary edge $f \in E \setminus \{e, g, h'\}$. Note that the pair of edges $\{g, f\}$ does not induce a coordinated subgraph of type $K(2, 1)$. Note also that, for an arbitrary edge $f \in E \setminus \{e, g, h, h'\}$, the triple of edges $\{g, h, f\}$ does not induce an interesting garland. Thus, we have built

$$|E| - 3 + |E| - 4 = 2|E| - 7$$

necessary subsets for the edge g. By analogy, we can build $2|E| - 7$ necessary subsets for the edge f.

Let $\hat{h}$ be an arbitrary edge from $E \setminus \{e, g, h, h'\}$. Consider the set $\{h, h', \hat{h}\}$. If the set $\{e, h, h', \hat{h}\}$ induces an interesting garland, this garland must be of type $K(2, 1) \dot{\cup} K(2, 1)$; the same is true for the set $\{g, h, h', \hat{h}\}$. Since both the sets $\{\hat{h}, h\}$ and $\{\hat{h}, h'\}$ cannot induce a coordinated subgraph of type $K(2,1)$, at least one of the sets $\{e, h, h', \hat{h}\}$ and $\{g, h, h', \hat{h}\}$ does not induce an interesting garland. Therefore, we have built $2|E| - 6$ necessary subsets for the edge e or for the edge g, and the lemma is proved if $|E| > 6$.

In the case where the edges e, g, h, and h' form the first configuration shown in Fig. 4, the sets $\{e, h, h'\}$ and $\{g, h, h'\}$ do not induce interesting garlands; consequently, we have built $2|E| - 5$ necessary subsets for the edge e or for the edge g.

It remains to prove the lemma in the case where $|E| = 5$ and the edges e, g, h, and h' form the second configuration shown in Fig. 4. Note that there are at most eight interesting garlands whose edge aggregates are subsets of $\{g, f, h', h\}$: four one-edge garlands, two garlands of type $K(2,1)$, and possibly two garlands of type $K(2,1)\dot{\cup}K(1,1)$. So it suffices to prove that there are at most five interesting garlands that contain the edge $\hat{h}$.

Case 2.3.2.1. Assume that the edge $\hat{h}$ is incident neither to the vertex x nor the vertex y. In this case, the edge $\hat{h}$ is not in any coordinated subgraph of type $K(2,1)$ because, otherwise, either the edge e is in two coordinated subgraphs of type $K(2,1)$ or the edge e is in two coordinated subgraphs of type $K(2,1)$, both the cases lead to a contradiction. Therefore, the edge $\hat{h}$ is not in any two-edge garland. Moreover, since $\langle E\rangle$ does not contain a triangle, the edge $\hat{h}$ is not in any one-element garland except for the one-edge garland.

Assume that the edge $\hat{h}$ is in some garland G of cardinality more than one. In this case, the garland G must destroy some part. The vertex aggregate $V(G)$ has at most two common vertices with any part $V \neq V_j$. Thus, V_j is the only part that can be destroyed by G. Note also that the edge $\hat{h}$ is not incident to the vertex z because, otherwise, the edge e or g is in two Ξ_2-subgraphs. Therefore, either $e \in E(G)$ or $g \in E(G)$. Note also that this two edges cannot be in $E(G)$ simultaneously; thus, the edge $\hat{h}$ is in at most two garlands with cardinality more than one. Therefore, in this case, the edge $\hat{h}$ is in at most three garlands, and the lemma is proved in this case.

Case 2.3.2.2. Now assume that the edge $\hat{h}$ is incident either to the vertex x or the vertex y. Without loss of generality, assume that $\hat{h}$ is incident to x. Note that, in this case, the edge aggregate that contains the edge $\hat{h}$ can be one of the following four sets: $\{\hat{h}\}$, $\{\hat{h}, h\}$, $\{\hat{h}, g, h'\}$, and $\{\hat{h}, h, g, h'\}$, and the lemma is proved in this case.

Case 2.4. It remains to consider the case where neither the edge e nor the edge g are in any coordinated subgraph of type $K(2,1)$. In this case, neither the set $\{e, f\}$ nor the set $\{g, f\}$ induce an interesting garland for any edge $f \in E \setminus \{e, g\}$, so we have built $|E| - 2$ necessary subsets for the edges e and g.

Since $|E \setminus \{e, g\}| \geqslant 3$, consider three arbitrary edges from this set and denote them by h_1, h_2 and h_3. Define $H = \{h_1, h_2, h_3\}$. Note that there exist exactly three two-element subsets $H' \subseteq H$. If, for any two-element subset $H' \subset H$, the set $H' \cup \{e\}$ does not induce an interesting garland, then we have built $|E|+1$ necessary subsets for the edge e. So it suffices to consider the case where, for some two-element subset $H' \subset H$, the set $H' \cup \{e\}$ induces an interesting garland. Without loss of generality, $H' = \{h_1, h_2\}$. Since $\langle E\rangle$ does not contain a triangle and the edge e is not in any coordinated subgraph of type $K(2,1)S$, an interesting three-edge garland must be of type $K(2,1)\dot{\cup}K(1,1)$ or $K(1,1,1)$.

Case 2.4.1. The set $H' \cup \{e\}$ induces an interesting garland of type $K(1,1,1)$. Note that, in this case, the set $\{g, h_1, h_2\}$ does not induce an interesting garland. Note also that at least one of the sets $\{g, h_3, h_1\}$ and $\{g, h_3, h_2\}$ does not induce an interesting garland. Therefore, we have built $|E|$ necessary subsets for the edge g, and, in this case, the lemma is proved.

Case 2.4.2. The set $H' \cup \{e\}$ induces an interesting garland of type $K(2,1)\dot{\cup}K(1,1)$. Note that, in this case, sets $\{g, h_3, h_1\}$ and $\{g, h_3, h_2\}$ does not induce interesting garlands. Therefore, we have built $|E|$ necessary subsets for the edge g and the lemma is proved. □

Lemma 12. *Assume that each active part of $K(v)$ contains at least three vertices. If an edge $e \in E$ is in $k \leqslant |E| - 2$ different Ξ_2-subgraphs, then*

(1) *the edge e is in at most $2^{|E|-k-1}$ interesting garlands;*

(2) *if $\langle E \rangle$ does not contain a coordinated subgraph of type $K(|E| - k, 1)$, then the edge e is in at most $2^{|E|-k-1} - |E| + k + 2$ interesting garlands.*

P r o o f. Let edges $h_1, h_2, \dots, h_k$ be such that the pair of edges $\{h_i, e\}$ induces a Ξ_2-subgraph for all $i = 1, 2, \dots, k$. Define $E' = E \setminus \{h_1, h_2, \dots, h_k\}$.

Let F be an edge aggregate of some interesting garland, and let $e \in F$. Note that $h_i \notin F$ for all $i = 1, 2, \dots, k$. Therefore, $F \subseteq E'$.

If E' does not induce a coordinated subgraph of type $K(|E| - k, 1)$, then, by Lemma 3, there exist at most $2^{|E|-k-1}$ interesting garlands whose edge aggregates are subsets of E'. Note that among them there are $|E| - k - 1$ one-edge garlands that do not contain the edge e. Consequently, the edge e is in at most

$$2^{|E|-k-1} + 1 - (|E| - k - 1) = 2^{|E|-k-1} - |E| + k + 2$$

interesting garland, and the lemma is proved. □

Lemma 13. *Assume that $|E| \geqslant 6$ and the subgraph $\langle E \rangle$ contains at least two Ξ_2-subgraphs and does not contain a coordinated subgraph of type $K(|E| - 2)$.*

(1) *If there exists an edge, which is in two Ξ_2-subgraphs, then the number of interesting garlands does not exceed $2^{|E|-2} + 2^{|E|-3} - |E| + 5$.*

(2) *If there exist two Ξ_2-subgraphs without common edges, then the number of interesting garlands does not exceed $2^{|E|-1} - 3|E| + 6$.*

P r o o f. Prove the first statement of the lemma. Assume that an edge e is in two Ξ_2-subgraphs. Since $\langle E \rangle$ does not contain a coordinated subgraph of type $K(|E| - 2)$, by Lemma 12, the edge e is in at most $2^{|E|-3} - |E| + 4$ interesting garlands. By Lemma 3, there are at most $2^{|E|-2} + 1$ interesting garlands whose edge aggregates are subsets of $E \setminus \{e\}$. Consequently, there are at most

$$2^{|E|-3} - |E| + 4 + 2^{|E|-2} + 1 = 2^{|E|-2} + 2^{|E|-3} - |E| + 5$$

interesting garlands.

Now prove the second statement of the lemma. Assume that there are two Ξ_2-subgraphs without common edges. Let edges e_1 and e_2 induce a Ξ_2-subgraph. By Lemma 9, the edge aggregate of any garland is a subset of $E \setminus \{e_1\}$ or $E \setminus \{e_2\}$. By Lemma 11, there are at most

$$2^{|E|-2} - (|E| - 1) + 1 = 2^{|E|-2} - |E| + 2$$

interesting garlands whose edge aggregates are subsets of $E \setminus \{e_1\}$ and there are at most $2^{|E|-2} - |E| + 2$ interesting garlands whose edge aggregates are subsets of $E \setminus \{e_2\}$. Note that $|E| - 2$ one-edge garlands were counted twice, so the number of interesting garlands does not exceed

$$2(2^{|E|-2} - |E| + 2) - (|E| - 2) = 2^{|E|-1} - 3|E| + 6,$$

and the lemma is proved. □

Lemma 14. *Assume that each active part of $K(v)$ contains at least three vertices and $\langle E\rangle$ contains a triangle. If an edge e is in the triangle from $\langle E\rangle$ and is in k different Ξ_2-subgraphs, then e is in at most $2^{|E|-k-2} - |E| + k + 3$ interesting garlands.*

P r o o f. Let Δ be the edge set of a triangle from $\langle E\rangle$ that contains the edge e, and let edges $h_1, h_2, \ldots, h_k$ be such that the pairs of edges $\{e, h_i\}$ induce Ξ_2-subgraphs for all $i = 1, 2, \ldots k$. Define $E' = E \setminus (\Delta \cup \{h_1, h_2, \ldots, h_k\})$.

Let G be an arbitrary garland such that $e \in E(G)$, and let $F = E(G)$. Note that $h_i \notin F$ for all $i = 1, 2, \ldots, k$. Define $F' = F \cap E'$. Note also that either $F \cap \Delta = \{e\}$ or $F \cap \Delta = \Delta$.

If $F \cap \Delta = \{e\}$, then F' can be chosen in $2^{|E|-k-3}$ ways. If $F \cap \Delta = \Delta$, then $|F'| > 1$ and, consequently, F' can be chosen in

$$2^{|E|-k-3} - (|E| - k - 3)$$

ways.

Thus, the edge e is in at most

$$2^{|E|-k-3} + 2^{|E|-k-3} - (|E| - k - 3) = 2^{|E|-k-2} - |E| + k + 3$$

interesting garlands. □

Lemma 15. *Assume that each part of $K(v)$ contains at least three vertices and $\langle E\rangle$ contains two triangles with a common edge e. If the edge e is in k different Ξ_2-subgraphs, then the edge e is in at most $2^{|E|-k-3} - 3|E| + 3k + 15$ interesting garlands.*

P r o o f. Let Δ_1 and Δ_2 be the edge sets of triangles from $\langle E\rangle$ that contain the edge e. Let edges $h_1, h_2, \ldots, h_k$ be such that the pair of edges $\{e, h_i\}$ induces a Ξ_2-subgraph for all $i = 1, 2, \ldots, k$. Define $E' = E \setminus (\Delta_1 \cup \Delta_2 \cup \{h_1, h_2, \ldots, h_k\})$ and $E_\Delta = \Delta_1 = \Delta_2$.

Let G be an arbitrary garland such that $e \in E(G)$. Define $F = E(G)$ and note that $h_i \notin F$ for all $i = 1, 2, \ldots, k$. Define $F' = F \cap E'$ and $F_\Delta = F \cap E_\Delta$. Note also that either $F_\Delta = \{e\}$, or $F_\Delta = \Delta_1$, or $F_\Delta = \Delta_2$, or $F_\Delta = \Delta$.

Consider an arbitrary $F' \subseteq E'$ and count the number of interesting garlands G such that $e \in E(G)$ and $F' \subseteq E(G)$. To do this, we estimate the number of ways to choose a subset $F_\Delta \subseteq \Delta$ such that $F' \cup F_\Delta$ induces an interesting garland. If $|F'| \neq 1$, then F_Δ can be chosen if four ways; if $|F'| = 1$, then $F\Delta \neq \Delta_1$ and $F\Delta \neq \Delta_2$. Let $f \in F'$. Note that if the set $\{f, e\}$ induces an interesting garland G', then G' is of type $K(2, 1)$ and $\{f\} \cup \Delta$ does not induce an interesting garland. Therefore, if $|F'| = 1$, then F_Δ can be chosen in at most one way.

Thus, the edge e is in at most

$$4 \cdot (2^{|E|-k-5} - (|E| - k - 5)) + (|E| - k - 5) = 2^{|E|-k-3} - 3|E| + 3k + 15$$

interesting garlands. □

5. Upper bound for the invariant pt when $|E|$ is small

The main goal of this section is to prove some upper bounds for the number of interesting garlands in the case where E contains relatively small number of elements.

Lemma 16. *Assume that each active part of $K(v)$ contains at least three vertices, $\langle E\rangle$ contains a triangle, does not contain a coordinated subgraph of type $K(4,1)$, and $|E| = 8$. Denote the edge set of the triangle by Δ. If $\langle E \setminus \Delta\rangle$ contains at least two Ξ_2-subgraphs, then the number of interesting garlands does not exceed* 46.

P r o o f. Let X be the number of interesting garlands whose edge aggregates are subsets of E', and let Y be the number of nonempty subsets of E' continuable outside E'. Let k be the number of edges $h \in E\setminus\Delta$ for which there exists an edge h' such that h and h' induce a coordinated subgraph of type $K(2,1)$.

Case 1. Assume that there exist three edges $e, f_1, f_2 \in E \setminus \Delta$ such that the pairs of edges $\{e, f_1\}$ and $\{e, f_2\}$ induce Ξ_2-subgraphs.

In this case, any continuable set can contain neither the set $\{e, f_1\}$ nor $\{e, f_2\}$. Consequently,

$$Y \leqslant 2^5 - 1 - 2\cdot 2^3 + 2^2 = 32 - 1 - 16 + 4 = 1.$$

By Lemma 3, there are at most $2^3 + 1 = 9$ interesting garlands whose edge aggregates are subsets of $E' \setminus \{e\}$.

Case 1.1. Assume that $E' \setminus \{f_1, f_2\}$ induces a coordinated subgraph of type $K(3,1)$. Denote this subgraph by H. In this case, the edge e is in exactly four interesting garlands whose edge aggregates are subsets of E'. Consequently, $X \leqslant 9 + 4 = 13$.

If there exists an edge from the triangle adjacent to all edges from H, then either $\langle E\rangle$ contains a coordinated subgraph of type $K(4,1)$, which contradicts the lemma hypothesis, or any edge from H does not induce a coordinated subgraph of type $K(2,1)$ with any edge from the triangle. Thus, $k \leqslant 2$. By Lemma 7, the number of interesting garlands does not exceed

$$2\cdot 13 + 2\cdot 19 - 5\cdot 8 + 19 + 2 \leqslant 26 + 38 - 40 + 19 + 2 = 45.$$

Case 1.2. Define $E'' = E' \setminus \{f_1, f_2\}$ and assume that E'' does not induce a coordinated subgraph of type $K(3,1)$. In this case, there exist at most 5 interesting garlands whose edge aggregates are subsets of E'', and the edge e is not in two (one-edge) of them. Consequently, the edge e is in at most three interesting garlands whose edge aggregates are subsets of E'. Therefore, $X \leqslant 9 + 3 = 12$, $k \leqslant 5$, and, using Lemma 7, one can deduce that the number of interesting garlands does not exceed

$$2\cdot 12 + 2\cdot 19 - 5\cdot 8 + 19 + 5 = 24 + 38 - 40 + 24 = 46.$$

Case 2. Now consider the case where $\langle E\rangle$ contains two Ξ_2-subgraph without common edges. Let the pairs of edges $\{e_1, e_2\}$ and $\{g_1, g_2\}$ induce Ξ_2-subgraphs.

Estimate the number of nonempty continuable sets. Let $F \subseteq E'$ be a nonempty continuable set, $F_e = F \cap \{e_1, e_2\}$, $F_g = F \cap \{g_1, g_2\}$, and let $F' = F \setminus \{e_1, e_2, g_1, g_2\}$. Note that $F = F_e \cup F_g \cup F'$. To obtain an upper bound, it suffices to count the number of ways to choose the sets F_e, F_g, and F'. Note that $F' \subseteq E' \setminus \{e_1, e_2, g_1, g_2\}$ and then it can be chosen in

$$2^{|E'\setminus\{e_1,e_2,g_1,g_2\}|} = 2^{5-4} = 2$$

ways. The set F_e can be either empty or one-element because, by Lemma 9, the edges e_1 and e_2 cannot be in the edge aggregate of any garland; so F_e can be chosen in three ways. By analogy, the set F_g can also be chosen in three ways. Excluding the empty set, one can conclude that $Y \leqslant 2\cdot 3\cdot 3 - 1 = 17$.

Note that the set $E' \setminus \{e_1\}$ contains a Ξ_2-subgraph and by Lemma 10, there are at most

$$2^{|E'\setminus\{e_1\}|-1} = 2^{4-1} = 8$$

interesting garlands whose edge aggregates are subsets of $E' \setminus \{e_1\}$. Note that three one-edge of them do not contain the edge e_2; consequently, e_2 is in at most 5 interesting garlands whose edge aggregates are subsets of E'. By analogy, there are at most 8 interesting garlands whose edge aggregates are subsets of $E' \setminus \{e_2\}$. Therefore, $X \leqslant 8 + 5 = 13$.

Thus, by Lemma 7, the number of interesting garlands does not exceed

$$2 \cdot 13 + 2 \cdot 17 - 5 \cdot 8 + 19 + 5 = 26 + 34 - 40 + 19 + 5 = 44,$$

and the lemma is proved. □

Lemma 17. *Assume that each part of $K(v)$ contains at least three vertices and $\langle E\rangle$ does not contain a coordinated subgraph of type $K(5,1)$. If $|E| = 7$ and $\langle E\rangle$ contains at least three Ξ_2-subgraphs, then the number of interesting garlands does not exceed* 41.

P r o o f. Assume that there exists an edge $e \in E$ that is in three Ξ_2-subgraphs. By Lemma 3, there exist at most $2^5 + 1 = 33$ interesting garlands whose edge aggregates are subsets of $E \setminus \{e\}$. By Lemma 12, one can deduce that e is in at most $2^{7-4} = 8$ interesting garlands. Therefore, there exist at most $33 + 8 = 41$ interesting garlands.

Now assume that there exists an edge $e \in E$ that is in exactly two Ξ_2-subgraphs. Since $\langle E\rangle$ does not contain a coordinated subgraph of type $K(5,1)$, by Lemma 12, one can deduce that e is in at most $2^{7-3} - 7 + 2 + 2 = 13$ interesting garlands. The graph $\langle E \setminus \{e\}\rangle$ contains a Ξ_2-subgraph and does not contain a coordinated subgraph of type $K(5,1)$. Thus, by Lemma 13, there exist at most $2^5 - 6 + 1 = 28$ interesting garlands whose edge aggregates are subsets of $E \setminus \{e\}$. Therefore, there exist at most $13 + 28 = 41$ interesting garlands.

It remains to consider the case when every edge is in at most one Ξ_2-subgraph. Let pairs of edges $\{e_1, e_2\}, \{f_1, f_2\}$, and $\{g_1, g_2\}$ induce three different Ξ_2-subgraphs. Denote the single left edge by h. By Lemma 13, there exist at most $2^5 - 3 \cdot 6 + 6 = 20$ interesting garlands whose edge aggregates are subsets of $E \setminus \{e_1\}$.

Estimate the number of interesting garlands that contain the edge e_1. Let H be the edge aggregate of such a garland. Define $H' = H \cap \{h\}, H_f = H \cap \{f_1, f_2\}$, and $H_g = H \cap \{g_1, g_2\}$ and note that $H = \{e\} \cup H' \cup H_f \cup H_G$. Note also that H' is empty or equal to $\{h\}$. The set H_f can be chosen in three ways (because the edges f_1 and f_2 cannot be in the same garland). By analogy, the set H_g can be chosen in three ways. Therefore, the set H can be chosen in at most $2 \cdot 3 \cdot 3 = 18$ ways. Therefore, e is in at most 18 interesting garlands, the total number of interesting garlands does not exceed $20 + 16 = 36$, and the lemma is proved. □

Lemma 18. *Assume that each part of $K(v)$ contains at least three vertices, $|E| = 7$, and $\langle E\rangle$ does not contain a coordinated subgraph of type $K(4,1)$. If every edge is in at most three Ξ_2-subgraphs and $\langle E\rangle$ contains at least four Ξ_2-subgraphs, then the number of interesting garlands does not exceed* 34.

P r o o f. Assume that any edge is not in two Ξ_2-subgraphs. Then

$$\xi_2 \leqslant \frac{|E|}{2} = \frac{7}{2} < 4,$$

a contradiction. Consequently, there exists an edge that is in at least two Ξ_2-subgraphs.

Case 1. Assume that there exist an edge e that is in three Ξ_2-subgraphs. By Lemma 12, there exist at most $2^{6-3-1}+1-3=6$ interesting garlands such that e is in them. Also, by Lemma 11, there are at most 28 interesting garlands whose edge aggregates are subsets of $E\setminus\{e\}$. Therefore, there exist at most $28+6=34$ interesting garlands.

Case 2. Assume that every edge is in at most two Ξ_2-subgraphs. Consider an edge e that is in exactly two Ξ_2-subgraphs. Let edges f_1 and f_2 be such that the pairs of edges $\{e,f_1\}$ and $\{e,f_2\}$ induce Ξ_2-subgraphs.

Case 2.1. Assume that $E\setminus\{f_1,f_2\}$ contains a Ξ_2-subgraph. By Lemma 11, there exists at most $2^4-5+1=12$ interesting garlands whose edge aggregates are subsets of $E\setminus\{f_1,f_2\}$. Note that four one-edge garlands from them do not contain the edge e. Therefore, e is at most in $12-4=8$ interesting garlands. The set $E\setminus\{e\}$ contains at least two Ξ_2-subgraphs and, by Lemma 13, there exist at most

$$\max(2^{6-2}+2^{6-3}-6+5,2^{6-1}-3\cdot 6+6)=\max(2^4+2^3-1,2^5-18+6)=\max(23,20)=23$$

interesting garlands whose edge aggregates are subsets of $E\setminus\{e\}$.

Case 2.2. Assume that $E\setminus\{f_1,f_2\}$ does not contain a Ξ_2-subgraph. By Lemma 12, the edge e is in at most 13 interesting garlands.

Case 2.2.1. Assume that $E\setminus\{e\}$ contains two Ξ_2-subgraphs without common edges. In this case, by Lemma 13, there exist at most $2^{6-1}-3\cdot 6+6=20$ interesting garlands whose edge aggregates are subsets of $E\setminus\{e\}$. Therefore, there exists at most $20+13=33$ interesting garlands.

Case 2.2.2. It remains to consider the case when there exists an edge $h\in E\setminus\{e\}$ that is in two Ξ_2-subgraphs. Since every Ξ_2-subgraph contains f_1 or f_2 (otherwise, $E\setminus\{f_1,f_2\}$ contains a Ξ_2-subgraph), the pairs of edges $\{e,f_1\}$, $\{e,f_2\}$, $\{h,f_1\}$, and $\{h,f_2\}$ induce Ξ_2-subgraphs. This means that the edge aggregate of any garland is a subset of $E\setminus\{f_1,f_2\}$ or $E\setminus\{e,h\}$. Applying Lemma 3 to each of them, one can conclude that there exist at most 17 interesting garlands whose edge aggregates are subsets of each of them. Since three one-edge garlands were counted twice, the total number of interesting garlands does not exceed $17+17-3=31$, and the lemma is proved. □

6. Proof of Theorem 1

The main goal of this section is to prove that a graph $K(n_1,n_2,n_3)$ is chromatically unique if

$$n_1\geqslant n_2\geqslant n_3\geqslant 2,\quad n_1-n_3=5\quad\text{and}\quad n_1+n_2+n_3\equiv 2\pmod 3.$$

The bottom levels of the lattice $NPL(n,3)$ in the case $n\equiv 2\pmod 3$ are shown in Fig. 5. The label above the covering relation shows how the number of edges changes, and the label under the covering relation shows how the invariant pt changes. To prove the theorem, it suffices to verify that the graph $K(q+3,q+1,q-2)$ is chromatically unique for $q\geqslant 4$ and the graph $K(q+4,q-1,q-1)$ is chromatically unique for $q\geqslant 3$.

Proposition 1. *The graph $K(q+3,q+1,q-2)$ is chromatically unique for $q\geqslant 4$.*

P r o o f. By contradiction, assume that the graph $K(u)=K(q+3,q+1,q-2)$ is not chromatically unique. This means that there exists a graph H such that the graphs $K(u)$ and H are chromatically equivalent. Let H be a graph obtained from the graph $K(v)$ by removing some

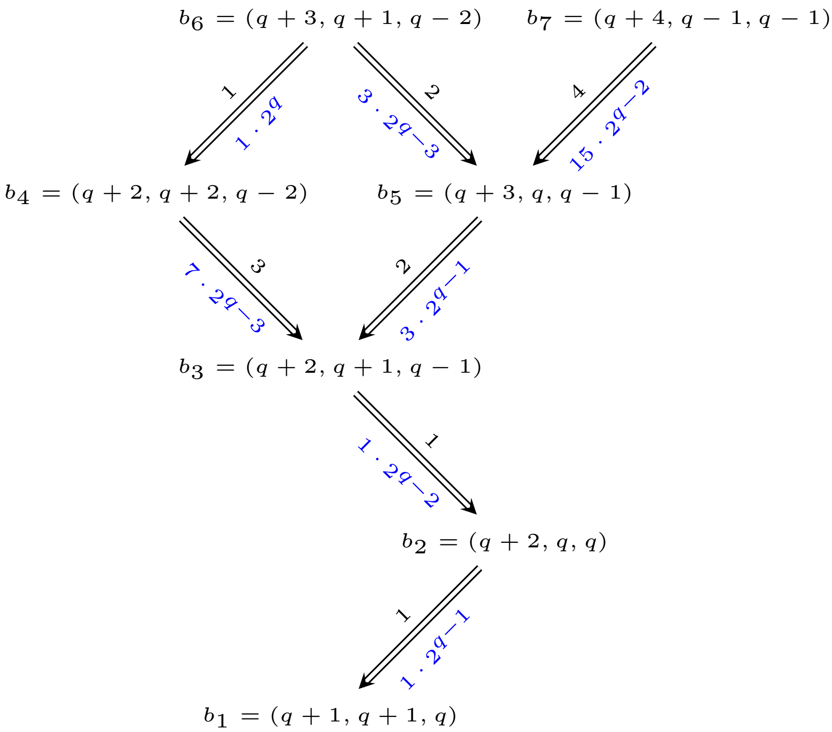


Figure 5. The bottom levels of the lattice $NPL(n,3)$ in the case $n \equiv 2 \pmod 3$

set of edges E. Consider the graph $K(v)$. Note that the cases $K(v) = K(q+2,q+2,q-2)$, $K(v) = K(q+3,q,q-1)$, and $K(v) = K(q+2,q+1,q-1)$ are impossible by Lemma 1.

Case 1. Assume that $K(v) = K(q+2,q,q)$ and note that, in this case, $|E| = 5$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta(K(u), K(v)) = 2^q + 7 \cdot 2^{q-3} + 2^{q-2} = 17 \cdot 2^{q-3}.$$

By Lemma 2, one can deduce that

$$17 \cdot 2^{q-3} \leqslant 2^5 - 1 = 31,$$

which implies $q \leqslant 3$, a contradiction.

Case 2. Assume that $K(v) = K(q+1,q+1,q)$. Note that, in this case, $|E| = 6$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta(K(u), K(v)) = 2^q + 7 \cdot 2^{q-3} + 2^{q-2} + 2^{q-1} = 21 \cdot 2^{q-3}.$$

By Lemma 2, one can deduce that

$$21 \cdot 2^{q-3} \leqslant 2^6 - 1 = 63,$$

which implies $q \leqslant 4$; so it suffices to check the case $q = 4$. In this case, $K(v) = K(5,5,4)$ and $\Delta pt(H, K(v)) = 42$. Note that, in this case, E does not induce a coordinated subgraph of type $K(5,1)$ because such a subgraph is a one-element garland that destroys one part, a contradiction. Thus, by Lemma 3, one can obtain that there are at most $2^5 + 1 = 33$ interesting garlands, which contradicts $\Delta pt(H, K(v)) = 42$, and the proposition is proved. $\square$

To prove the chromatic uniqueness of the graph $K(q+4, q-1, q-1)$, we need the following two statements.

Lemma 19. *Let E be a subset of edges of some complete tripartite graph $K(v)$. If any two triangles from $\langle E\rangle$ have a common edge, then there exists an edge that is in all triangles from $\langle E\rangle$.*

Before proving this lemma, note that the condition on the graph $K(v)$ to be tripartite is necessary. Four triangles, any two of which have a common edge but no edge is common for all of them, are shown in Fig. 6.

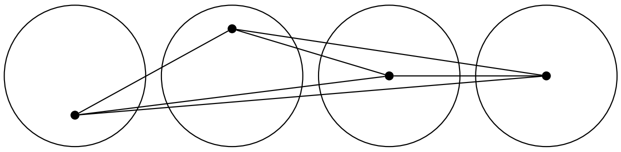


Figure 6. Four triangles, any two of which have a common edge but no edge is common for all of them.

P r o o f. Let Δ_1 and Δ_2 be two triangles from $\langle E\rangle$ that have a common edge. Let x, y_1, and z be vertices of the triangle Δ_1 and let x, y_2, and z be vertices of the triangle Δ_2. Let us prove that if a triangle Δ from $\langle E\rangle$ have a common edge with the triangles Δ_1 and Δ_2, then $xz \in \Delta$.

By contradiction, assume that $xz \notin \Delta$. The triangles Δ_1 and Δ have a common edge, without loss of generality, let it be the edge y_1z. Denote by x' the third vertex of Δ_1 and note that the vertices x and x' are in the same part of $K(v)$ because $K(v)$ is a tripartite graph. Note that, in this case, the edge set of Δ_1 is $\{x'y_1, x'z_1, y_1z_2\}$ and it has an empty intersection with the edge set of Δ_2 (see Fig. 7). □

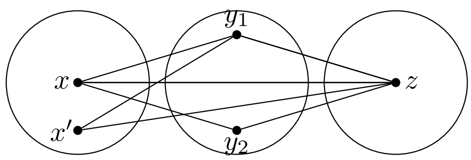


Figure 7. Three triangles in the tripartite graph that do not have a common edge

Proposition 2. *If a graph H is obtained by removing the edge set E from the graph $K(4, 4, 3)$, then the graphs H and $K(7, 2, 2)$ are not chromatically equivalent.*

P r o o f. The graph $K(7, 2, 2)$ has $7 \cdot 2 + 7 \cdot 2 + 2 \cdot 2 = 32$ edges and the graph $K(4, 4, 3)$ has $4 \cdot 4 + 4 \cdot 3 + 4 \cdot 3 = 40$ edges; consequently, $|E| = 40 - 32 = 8$.

Calculate the difference of the invariant I_3:

$$I_3(K(7,2,2)) = 7 \cdot 2 \cdot 2 = 28,$$
$$I_3(K(4,4,3)) = 4 \cdot 4 \cdot 3 = 48,$$
$$20 = \Delta I_3(K(4,4,3), K(7,2,2)) = \Delta I_3(K(4,4,3), H) = \xi_1 - \xi_2 - 2\xi_3,$$

$$\xi_1 = 3e_{12} + 4e_{13} + 4e_{23} = 4|E| - e_{12},$$
$$20 = 32 - e_{12} - \xi_2 - 2\xi_3,$$
$$\xi_2 + 2\xi_3 + e_{12} = 12.$$

Note that every edge is in at most four Ξ_2-subgraphs.

The inequality $e_{12} \leqslant 8$ implies $\xi_2 + 2\xi_3 \geqslant 4$. Note that $E \not\subseteq E_{ij}$ for all nonequal i and j from 1 to 3 because, otherwise, $\xi_2 = \xi_3 = 0$, a contradiction. Assume that $e_{ij} = |E| - 1$ for some i and j. In this case, $\xi_3 = 0$ and $\xi_2 \geqslant 5$. Denote the single edge from $E \setminus E_{ij}$ by f. Note that all Ξ_2-subgraphs must contain f; therefore, $\xi_2 \leqslant 4$, a contradiction.

Calculate the difference of the invariant pt:

$$pt(K(7,2,2)) = 2^6 + 2^1 + 2^1 - 3 = 65,$$
$$pt(K(4,4,3)) = 2^3 + 2^3 + 2^2 - 3 = 17,$$
$$\Delta pt(H, K(4,4,3)) = \Delta pt(K(7,2,2), K(4,4,3)) = 65 - 17 = 48.$$

Note that $\langle E \rangle$ does not contain a coordinated subgraph of type $K(4,1)$ because, otherwise, such a subgraph forms a one-element garland that destroys one part, a contradiction.

Case 1. Assume that $\xi_3 = 0$.

Case 1.1. Assume that $e_{12} = 6$. In this case, $\xi_2 = 6$.

Let edges f and g be not in E_{12}. Without loss of generality, assume that $f \in E_{23}$.

Note that any edge from E except for f and g is not adjacent to any vertex from the part V_3. Since $|V_3| = 3$, there exists a vertex in V_3 not incident to any edge from E. Consequently, any garland cannot destroy the part V_3.

By Lemma 3, there exist at most $2^5 + 1 = 33$ interesting garlands whose edge aggregates are subsets of E_{12}; so it suffices to prove that the number of interesting garlands G such that $f \in E(G)$ or $g \in E(G)$ does not exceed 12 because, in this case, the total number of interesting garlands does not exceed $33 + 12 = 45 < 48$.

Case 1.1.1. Assume that edges f and g induce a Ξ_2-subgraph. Then there exist no garland G such that $f \in E(G)$ and $g \in E(G)$. Note also that, since every edge is in at most four Ξ_2-subgraphs and every Ξ_2-subgraph contains f or g, each of the edges f and g is in at least three Ξ_2-subgraphs.

Note that if f is in a garland G such that $|E(G)| > 1$, then $|G| > 1$; hence, G must destroy some part. It cannot destroy the part V_3, so it must destroy V_1 or V_2; in both cases, $|E(G)| \geqslant 4$. Note that there exist at least two edges from E_{12} such that f induces a Ξ_2-subgraph with each of them. This means that there exist at most four edges $h \in E_{12}$ such that f and h can be in the edge aggregate of G. Note also that $E(G)$ must contain at least three edges from E_{12}, and these three edges can be chosen in $\binom{4}{3} = 4$ ways. Therefore, f is in at most five interesting garlands. By analogy, g is in at most five interesting garlands, and the proof is complete in this case.

Case 1.1.2. Assume that edges f and g induce a coordinated subgraph of type $K(2,1)$. Note that, in this case, each of the edges f and g is in exactly three different Ξ_2-subgraphs. Note that if f is in a garland G that contains more than one element, then G must destroy some part. Since it cannot destroy V_3, it must destroy V_1 or V_2; in both cases $|E(G)| \geqslant 4$.

Let $F \subseteq E_{12}$. Note that the following three statements are equivalent for the set F.

(1) The set $F \cup \{f\}$ induces an interesting garland.

(2) The set $F \cup \{g\}$ induces an interesting garland.

(3) The set $F \cup \{f, g\}$ induces an interesting garland.

The equivalence of these three statements follows from the following two observations. First, if f is in some garland G, then $\{f\}$ forms an E-subgraph from G; and the same statement holds for $\{g\}$ and $\{g, f\}$. Therefore, the sets $F \cup \{f\}$, $F \cup \{g\}$, and $F \cup \{f, g\}$ induce a garland simultaneously. Second, denote by V_f the vertex set of $\langle F \cup \{f\} \rangle$, by V_g the vertex set of $\langle F \cup \{g\} \rangle$, and by V_{gf} the vertex set of $\langle F \cup \{g, f\} \rangle$. Note that $V_f \cap (V_1 \cup V_2) = V_g \cap (V_1 \cup V_2) = V_{gf} \cap (V_1 \cup V_2)$. This means that the sets $F \cup \{f\}$, $F \cup \{g\}$, and $F \cup \{f, g\}$ destroy the parts V_1 or V_2 simultaneously. Consequently, all of them induce interesting garlands or none of them does.

Note also that F is chosen from a three-element set formed by edges $h \in E_{12}$ such that the edges h and f do not induce a Ξ_2-subgraph. Consequently, the set F can be chosen in at most one way. Therefore, there exist at most $3 \cdot 1 = 3$ interesting garlands G such that $f \in E(G)$ or $g \in E(G)$, and the prove is complete in this case.

Case 1.1.3. Assume that edges f and g do not induce a coordinated subgraph of type $K(2, 1)$ and do not induce a Ξ_2-subgraph.

Note that, in this case, each of the edges f and g is in exactly three Ξ_2-subgraphs. Note that if f is in a one-element garland G, then G is a one-edge garland. Note also that if f is in a garland G that has more than one element, then G must destroy V_1 or V_2, and then $|E(G)| \geqslant 4$. Since f is in three Ξ_2-subgraphs, the set $E(G) \setminus \{f\}$ is chosen from a five-element set of edges. Consequently, f is in at most

$$\binom{4}{3} + \binom{4}{4} = 4 + 1 = 5$$

interesting garlands that contain more than one element. Therefore, f is in at most six interesting garlands. By analogy, the edge g is in at most six interesting garlands. Therefore, the number of interesting garlands whose edge aggregates contain f or g does not exceed 12, and the lemma is proved in Case 1.1.

Case 1.2. Assume that $e_{12} \leqslant 5$. In this case, $\xi_2 \leqslant 5$.

Case 1.2.1. Assume that there exists an edge $f \in E$ that is in four Ξ_2-subgraphs. Thus, by Lemma 12, it is in at most $2^{8-4-1} - 8 + 6 = 6$ interesting garlands. Note that the set $E \setminus \{f\}$ contains at least three Ξ_2-subgraphs and then, by Lemma 17, there exist at most 41 interesting garlands whose edge aggregates are subsets of $E \setminus \{f\}$. Consequently, there are at most $6 + 41 = 47$ interesting garlands, a contradiction with $\Delta pt(H, K(v)) = 48$.

Case 1.2.2. Assume that there exists an edge $f \in E$ which is in exactly three Ξ_2-subgraphs. Thus, by Lemma 12, it is in at most $2^{8-3-1} - 8 + 5 = 13$ interesting garlands. Note that $E \setminus \{f\}$ contains at least four Ξ_2-subgraphs; so, by Lemma 18, there exist at most 34 interesting garlands whose edge aggregates are subsets of $E \setminus \{f\}$. Therefore, there exist at most $13 + 34 = 47$ interesting garlands, a contradiction with $\Delta pt(H, K(v)) = 48$.

Case 1.2.3. It remains to consider the case when every edge from E is in at most two Ξ_2-subgraphs. Note that, since every Ξ_2-subgraph contains two edges, $\xi_2 \leqslant 8$. If $e_{12} = 5$, then every Ξ_2-subgraph must contain at least one of three edges from $E' = E \setminus E_{12}$. Consequently, some edge from E' must be in three Ξ_2-subgraphs, a contradiction. Therefore, $e_{12} \leqslant 5$ and then $\xi_2 \geqslant 8$. Thus, $\xi_2 = 8$ and e_{12}. Note that, in this case, every edge is in exactly two Ξ_2-subgraphs. Since $e_{12} = 4$ and Ξ_2-subgraph cannot contain both edges from E_{12}, every Ξ_2-subgraph contain an edge from E_{12}.

Let F be the edge aggregate of some garland. Note that F does not contain edges of any Ξ_2-subgraph. Consequently, every Ξ_2-subgraph contains at least one edge from $E \setminus F$. Since every edge is in at most two Ξ_2-subgraphs, $8 = \xi_2 \leqslant 2|E \setminus F|$, which implies $|E \setminus F| \geqslant 4$, and then $|F| \leqslant 4$.

Assume that there exists a four-edge garland. Denote its edge aggregate by F and define $F' = E \setminus F$. Since $|F'| = 4$, there exists at most one four-edge garland whose edge aggregate is a subset of F'. Note that every Ξ_2-subgraph contains one edge from F and one edge from F'. Estimate the number of four-edge garlands whose edge aggregates contain an edge from F and an edge from F'. Assume that such a garland exists. Denote its edge aggregate by $\tilde{F}$. Define $F = \{f_1, f_2, f_3, f_4\}$ and $F' = \{f'_1, f'_2, f'_3, f'_4\}$. Let $f_1 \in \tilde{F}$ and $f'_1 \in \tilde{F}$. Without loss of generality, assume that the pairs of edges $\{f_1, f'_2\}, \{f_1, f'_3\}, \{f'_1, f_2\}$, and $\{f'_1, f_3\}$ induce Ξ_2-subgraphs. Then

$f_4 \in \tilde{F}$. This means that every edge from F is in at most one such four-edge garland and each such garland must contain exactly two edges from F; so the number of such garlands does not exceed $|F|/2 = 2$. Therefore, the number of four-edge garlands does not exceed 4.

Since $\langle E\rangle$ does not contain a triangle, an interesting three-edge garland must be of type $K(3,1)$ or $K(2,1)\dot{\cup}K(1,1)$. Since $e_{12} = 4$, the inequality $e_{ij} \leqslant 4$ holds for all i and j. Note that all three edges of a garland of type $K(3,1)$ must be in E_{ij} for some i and j. Since $\langle E\rangle$ does not contain a coordinated subgraph of type $K(4,1)$, any two garlands of type $K(3,1)$ cannot contain more than one common edge. Therefore, there exist at most two such garlands.

An interesting garland of type $K(2,1)\dot{\cup}K(1,1)$ must destroy some part of $K(v)$. Since it can destroy only a part that contains three vertices, it must destroy V_3. Consequently, each vertex of V_3 must be incident to some edge from this garland. Therefore, its edge aggregate must be a three-element subset of $E \setminus E_{12}$. Since $|E \setminus E_{12}| = 4$, such a subset can be chosen in $\binom{4}{3} = 4$ ways; so the number of such garlands does not exceed 4.

Thus, there exist 8 one-edge garlands, at most

$$\binom{8}{2} = 28$$

interesting two-edge garlands, at most $4 + 2 = 6$ interesting garlands, and at most 4 four-edge interesting garlands. Therefore, there exist at most

$$8 + 28 + 6 + 4 = 46 < 48$$

interesting garlands. The proof is complete in Case 1.

Case 2. Assume that $\xi_3 = 1$.

Case 2.1. Assume that $e_{12} = 6$. Then $\xi_2 = 4$ and $|E_{23}| = |E_{13}| = 1$. Denote the single edge from E_{23} by e and the single edge from E_{13} by f. Note that every Ξ_2-subgraph contains the edge e or f. If e is in three Ξ_2-subgraphs, then $\langle E\rangle$ contains a coordinated subgraph of type $K(4,1)$, a contradiction. By analogy, f is in at most two Ξ_2-subgraphs. This means that 7 edges from $|E|$ form a configuration shown in Fig. 8.

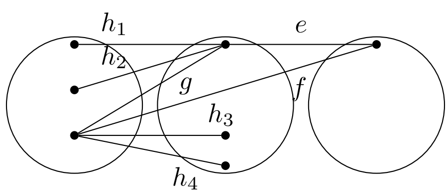


Figure 8. The case $e_{12} = 6$ and $\xi_3 = 1$

Note that any garland cannot destroy the part V_3.

Assume that e and f are in some garland G that contains more than one element. Note that $g \in E(G)$ and $h_1, h_2, h_3, h_4 \notin E(G)$ because, otherwise, $\langle E\rangle$ contains at least two triangles. Consequently, $|E(G)| \leqslant 4$ and $|V(G) \cap V_i| \leqslant 2$ for all $i = 1, 2, 3$. Therefore, G cannot destroy any part, a contradiction. Thus, if e and f are in a garland G, then $|G| = 1$. Note that there exists only one such garland, which is a triangle.

Note that e is not in any coordinated subgraph of type $K(2,1)$. This means that there are only two one-element garlands whose edge aggregates contain e: a triangle and a one-edge garland. Assume that e is in some garland G that contains more than one element and $f \notin E(G)$. Since G must destroy some part and cannot destroy V_3, it destroys V_1 or V_2; hence, $|E(G)| \geqslant 4$. Note that $f, g, h_1, h_2 \notin E(G)$, which implies $E(G) = E \setminus \{f, g, h_1, h_2\}$. Therefore, e is in at most one interesting garland that contains more than one element and does not contain f. By analogy, f is in at most one interesting garland that contains more than one element and does not contain e. Therefore, there exist at most 5 interesting garlands that contain e or f.

By Lemma 3, there exist at most 33 interesting garlands whose edge aggregates are subsets of E_{12}. Therefore, there exist at most $33 + 5 = 38$ interesting garlands, which contradicts $\Delta pt(H, K(v)) = 48$.

Case 2.2. Assume that $e_{12} \leqslant 5$. In this case, $\xi_2 \geqslant 5$. Note that any edge from the triangle cannot be in more than three Ξ_2-subgraphs.

Assume that some edge e from the triangle is in three Ξ_2-subgraphs. By Lemma 14, one can deduce that e is in at most $2^{8-2-3} - 8 + 3 + 3 = 6$ interesting garlands. Since $\langle E \setminus \{e\}\rangle$ contains at least three Ξ_2-subgraphs, by Lemma 17, there exist at most 41 interesting garlands whose edge aggregates are subsets of $E \setminus \{e\}$. Therefore, there exist at most $41 + 6 = 47$ interesting garlands, a contradiction.

Assume that some edge e from the triangle is in two Ξ_2-subgraphs. By Lemma 14, one can deduce that e is in at most $2^{8-2-2} - 8 + 2 + 3 = 13$ interesting garlands. Since $\langle E \setminus \{e\}\rangle$ contains at least three Ξ_2-subgraphs, by Lemma 18, there exist at most 34 interesting garlands whose edge aggregates are subsets of $E \setminus \{e\}$. Therefore, there exist at most $13 + 34 = 47$ interesting garlands, a contradiction.

By Lemma 16, there exists at most one Ξ_2-subgraph such that both its edges are not in the triangle. Taking into account that each edge of the triangle is in at most one Ξ_2-subgraph, one can deduce that $\xi_2 \leqslant 3 + 1 = 4 < 5$, a contradiction.

Case 3. Consider the case $\xi_3 \geqslant 2$.

If $\langle E\rangle$ contains two triangles without common edges, then, by Lemma 8, the number of interesting garlands does not exceed $2^6 - 12 \cdot 8 + 58 = 26 < 48$, a contradiction. Consequently, any two triangles from $\langle E\rangle$ have a common edge and, by Lemma 19, there exists an edge e that is in all triangles. This implies $\xi_3 \leqslant 3$. Define $E' = E \setminus \{e\}$.

Note that there exist at most two Ξ_2-subgraph that do not contain the edge e. Indeed, any such Ξ_2-subgraph does not contain any edge from any triangle. Therefore, if there are at least three such subgraphs, then some two of them do not contain edges from the same triangle and, in this case, by Lemma 16, the number of interesting garlands does not exceed 46.

Let e be in k different Ξ_2-subgraphs. Note that $k \leqslant 2$. Since e is in ξ_3 triangles, there exist ξ_3 pairs of edges $\{h, f\}$ such that h and f induce a Ξ_2-subgraph in $\langle E'\rangle$ (such pairs can be generated by removing the edge e from a triangle from $\langle E\rangle$). Therefore, $\langle E'\rangle$ contains at least

$$\xi_2 + \xi_3 - k = 12 - e_{12} - 2\xi_3 + \xi_3 - 2 = 12 - k - e_{12} - \xi_3 = \xi_2'$$

different Ξ_2-subgraphs. Note that if $e_{12} = 5$, then $\xi_3 \leqslant 2$; if $e_{12} \leqslant 4$, then $\xi_3 \leqslant 3$. In both cases, $e_{12} + \xi_3 \leqslant 7$ and this implies $\xi_2' \geqslant 12 - k - 7 = 5 - k$.

Case 3.1. Assume that $k = 2$. By Lemma 15, the edge e is in at most $2^{8-5} - 3 \cdot 8 + 6 + 15 = 5$ interesting garlands. The graph $\langle E'\rangle$ contains at least 3 different Ξ_2-subgraphs, therefore, by Lemma 17, there exist at most 41 interesting garlands whose edge aggregates are subsets of E'. Thus, there exist at most $41 + 5 = 46 < 48$ interesting garlands, a contradiction.

Case 3.2. Assume that $k = 1$. By Lemma 15, the edge e is in at most $2^{8-4} - 3 \cdot 8 + 3 + 15 = 10$ interesting garlands. The subgraph $\langle E'\rangle$ contains at least four Ξ_2-subgraphs, therefore, by Lemma 18, there exist at most 34 interesting garlands whose edge aggregates are subsets of E'. Thus, the number of interesting garlands does not exceed $10 + 34 = 44 < 48$, a contradiction.

Case 3.3. It remains to consider the case when the edge e is not in any Ξ_2-subgraph. Note that, in this case, $\xi_2 \leqslant 2$.

If $e_{12} = 5$, then $\xi_3 = 2$ and $\xi_2 = 12 - 2 \cdot 2 - 5 = 3 > 2$, a contradiction.

If $e_{12} \leqslant 3$, then $\xi_3 \leqslant 3$ and $\xi_2 = 12 - (e_{12} + 2\xi_3) \geqslant 12 - 9 = 3 > 2$, a contradiction.

Thus, it suffices to consider the case when $e_{12} = 4$. In this case, $\xi_3 = 3$ and $\xi_2 = 2$. Consider the single edge that is not in any triangle. Note that this edge must be in two Ξ_2-subgraphs, which is impossible. □

Now we are ready to prove the chromatic uniqueness of the graph $K(q+4, q-1, q-1)$.

Proposition 3. *The graph $K(q+4, q-1, q-1)$ is chromatically unique for $q \geqslant 3$.*

P r o o f. By contradiction, assume that the graph $K(u) = K(q+4, q-1, q-1)$ is not chromatically unique. This means that there exists a graph H such that the graphs $K(u)$ and H are chromatically equivalent. Let H be a graph obtained from the graph $K(v)$ by removing some edge set E. Consider the graph $K(v)$. Note that the case $K(v) = K(q+3, q, q-1)$ is impossible by Lemma 1.

Case 1. Assume that $K(v) = K(q+3, q+1, q-2)$ and note that, in this case, $|E| = 2$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta pt(K(u), K(v)) = 15 \cdot 2^{q-2} - 3 \cdot 2^{q-3} = 27 \cdot 2^{q-3}.$$

By Lemma 2, one can deduce that $27 \cdot 2^{q-3} \leqslant 2^2 - 1 = 3$, which implies $q < 2$, a contradiction.

Case 2. Assume that $K(v) = K(q+2, q+2, q-2)$ and note that, in this case, $|E| = 3$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta pt(K(u), K(v)) = 15 \cdot 2^{q-2} - 3 \cdot 2^{q-3} + 2^q = 35 \cdot 2^{q-3}.$$

By Lemma 2, one can deduce that $33 \cdot 2^{q-3} \leqslant 2^3 - 1 = 7$, which implies $q < 2$, a contradiction.

Case 3. Assume that $K(v) = K(q+2, q+1, q-1)$ and note that, in this case, $|E| = 6$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta pt(K(u), K(v)) = 15 \cdot 2^{q-2} + 3 \cdot 2^{q-1} = 21 \cdot 2^{q-2}.$$

By Lemma 2, one can deduce that $21 \cdot 2^{q-2} \leqslant 2^6 - 1 = 63$, which implies $q \leqslant 3$, so it suffices to consider the case $q = 3$. In this case, $K(v) = K(5, 4, 2)$ and $\Delta(H, K(v)) = 42$. Calculate the difference of the invariant I_3:

$$\begin{gathered}
\Delta I_3(K(v), K(u)) = 6(q-1), \\
\Delta I_3(K(v), H) = \xi_1 - \xi_2 - 2\xi_3, \\
\xi_1 = (q-1)e_{12} + (q+1)e_{13} + (q+2)e_{23} = (q-1)|E| + 2e_{13} + 3e_{23}, \\
\xi_2 + 2\xi_3 = 2e_{13} + 3e_{23}.
\end{gathered}$$

If $e_{13} = e_{23} = 0$, then $E = E_{12}$ and each active part of $K(v)$ contains at least four vertices. Therefore, by Lemma 3, there exist at most 33 interesting garlands, a contradiction. Thus, it suffices to consider the cases when $e_{13} > 0$ or $e_{23} > 0$. In both cases, $\xi_2 + 2\xi_3 \geqslant 2$.

Case 3.1. Assume that $\xi_3 > 0$ and edges $e_1, e_2, e_3 \in E$ induce a triangle. Consider an arbitrary garland G and define $F' = E(G) \cap \{e_1, e_2, e_3\}$. Since F' can be one of five sets (empty, one-element or equal to $\{e_1, e_2, e_3\}$), there exist at most $5 \cdot 2^{6-3} - 1 = 39 < 42$, a contradiction.

Case 3.2. Now assume that $\xi_3 = 0$. This implies $\xi_2 \geqslant 2$.

Case 3.2.1. Assume that there exists an edge e that is in two Ξ_2-subgraphs. Let f and g be edges such that $\{e, g\}$ and $\{e, f\}$ induce Ξ_2-subgraphs. Let G be a garland. Define $E' = E(G) \cap \{e, g, f\}$. Since E' can be one of five sets (empty, three one-element sets, and $\{g, f\}$), the number of garlands does not exceed $5 \cdot 2^{6-3} = 40 < 42$, a contradiction.

Case 3.2.2. Now assume that there exist two Ξ_2-subgraphs without common edges. Let e_1, e_2, f_1, and f_2 be edges from E such that $\{e_1, e_2\}$ and $\{f_1, f_2\}$ induce Ξ_2-subgraphs. Let G be a garland. Define $E_e = E(G) \cap \{e_1, e_2\}$, $E_f = E(G) \cap \{f_1, f_2\}$, and $E' = E(G) \setminus \{e_1, e_2, f_1, f_2\}$. Note that $E(G) = E_e \cup E_f \cup E'$. Since E_e can be chosen in three ways, E_f can be chosen in three ways, and E' can be chosen in $2^2 = 4$ ways, the number of garlands does not exceed $3 \cdot 3 \cdot 4 = 36 < 42$, a contradiction.

Case 4. Assume that $K(v) = K(q+2, q, q)$ and note that, in this case, $|E| = 7$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta pt(K(u), K(v)) = 15 \cdot 2^{q-2} + 3 \cdot 2^{q-1} + 2^{q-2} = 22 \cdot 2^{q-2}.$$

By Lemma 2, one can deduce that $22 \cdot 2^{q-2} \leqslant 2^7 - 1 = 127$, which implies $q \leqslant 4$. So it suffices to consider the cases when $q = 4$ or $q = 3$. If $q = 4$, then $K(v) = K(6, 4, 4)$ and $\Delta pt(H, K(v)) = 88$. By Lemma 3, the number of interesting garlands does not exceed $2^6 + 1 = 65 < 88$, a contradiction. Thus, it suffices to consider the case $q = 3$. In this case, $K(v) = K(5, 3, 3)$ and $\Delta pt(H, K(v)) = 44$. Note that $\langle E \rangle$ does not contain a coordinated subgraph of type $K(5, 1)$.

Calculate the difference of the invariant I_3:

$$\begin{gathered}
\Delta I_3(K(v), K(u)) = 7q - 4, \\
\Delta I_3(K(v), H) = \xi_1 - \xi_2 - 2\xi_3, \\
\xi_1 = qe_{12} + qe_{13} + (q+2)e_{23} = q|E| + 2e_{23}, \\
\xi_2 + 2\xi_3 = 2e_{23} + 4.
\end{gathered}$$

If $\xi_3 > 0$, then, by Lemma 6, the number of interesting garlands does not exceed

$$2^5 + 2^4 - 3 \cdot 7 + 13 = 32 + 16 - 21 + 13 = 40 < 44,$$

a contradiction. Therefore, $\xi_3 = 0$, which implies $\xi_2 = 2e_{23} + 4 \geqslant 4$. By Lemma 17, the number of interesting garlands does not exceed 41, a contradiction.

Case 5. Assume that $K(v) = K(q+1, q+1, q)$ and note that, in this case, $|E| = 8$. Calculate the difference of the invariant pt:

$$\Delta pt(H, K(v)) = \Delta pt(K(u), K(v)) = 15 \cdot 2^{q-2} + 3 \cdot 2^{q-1} + 2^{q-2} + 2^{q-1} = 24 \cdot 2^{q-2}.$$

Using Lemma 2, one can deduce that $24 \cdot 2^{q-1} \leqslant 2^8 - 1 = 255$, which implies $q \leqslant 6$; so it suffices to consider the cases when $q = 5$, $q = 4$, or $q = 3$. The case $q = 3$ is impossible by Proposition 2.

If $q = 5$, then $K(v) = K(6, 6, 5)$ and $\Delta pt(H, K(v)) = 192$. Since $\langle E \rangle$ dos not contain a coordinated subgraph of type $K(8, 1)$, by Lemma 3, the number of interesting garlands does not exceed $2^7 + 1 = 129 < 192$, a contradiction.

It remains to prove the proposition in the case $q = 4$. In this case, $K(v) = K(5, 5, 4)$ and $\Delta pt(H, K(v)) = 96$.

Calculate the difference of the invariant I_3:

$$\begin{gathered}\Delta I_3(K(v), K(u)) = 8q - 4,\\ \Delta I_3(K(v), H) = \xi_1 - \xi_2 - 2\xi_3,\\ \xi_1 = qe_{12} + (q+1)e_{13} + (q+1)e_{23} = (q+1)|E| - e_{12},\\ 8q - 4 = 8(q+1) - e_{12} - \xi_2 - 2\xi_3,\\ 12 = \xi_2 + 2\xi_3 + e_{12}.\end{gathered}$$

If $\xi_3 > 0$, then, by Lemma 6, the number of interesting garlands does not exceed

$$2^6 + 2^5 - 3 \cdot 8 + 13 = 85 < 96,$$

a contradiction. Thus, $\xi_3 = 0$, which implies $\xi_2 + e_{12} = 12$. Note that $e_{12} \leqslant 8$. If $e_{12} = 8$, then $E = E_{12}$, which implies $\xi_2 = 0$, a contradiction. Thus, $e_{12} \leqslant 7$ and $\xi_2 \geqslant 5$.

If there exist an edge e that is in at least two Ξ_2-subgraphs, then, since $\langle E\rangle$ does not contain a coordinated subgraph of $K(6,1)$, by Lemma 12, the edge e is in at most $2^5 - 8 + 2 + 2 = 28$ interesting garlands. By Lemma 3, the number of interesting garlands whose edge are subsets of $E \setminus \{e\}$ does not exceed $2^6 + 1 = 65$, so the total number of interesting garlands does not exceed $65 + 28 = 93 < 96$, a contradiction. Therefore, every edge is in at most one Ξ_2-subgraphs, which implies $\xi_2 \leqslant 4$, a contradiction. □

The proof of the theorem follows from Propositions 1 and 3.

7. Conclusion

In this paper, the chromatic uniqueness of a complete tripartite graph $K(n_1, n_2, n_3)$ is proved for $n_1 \geqslant n_2 \geqslant 3 \geqslant 2$ and $n_1 - n_3 \leqslant 5$. Also, some properties of the number of partitions of the vertex set of a graph G into t independent sets are established. Both problems, the chromatic uniqueness, and properties of the invariant pt, are still challenging open problems.

Acknowledgements

The author is grateful to his scientific advisor prof. V.A. Baransky for constant attention and remarks.

REFERENCES

1. Asanov M. O., Baransky V. A., Rasin V. V. *Diskretnaya matematika: grafy, matroidy, algoritmy* [Discrete Mathematics: Graphs, Matroids, Algorithms]. Saint-Petersburg: “Lan”, 2010. 364 p. (in Russian)
2. Baransky V. A., Koroleva T. A. Chromatic uniqueness of certain complete tripartite graphs. *Izv. Ural. Gos. Univ. Mat. Mekh. Inform.*, 2010. Vol. 74, Suppl. 12. P. 5–26. (in Russian)
3. Baransky V. A., Koroleva T. A., Senchonok T. A. On the partition lattice of all integers. *Sib. Èlektron. Mat. Izv.*, 2016. Vol. 13. P. 744–753. DOI: 10.17377/semi.2016.13.060 (in Russian)
4. Baranskii V. A., Sen’chonok T. A. Chromatic uniqueness of elements of height ≤ 3 in lattices of complete multipartite graphs. *Proc. Steklov Inst. Math.*, 2012. Vol. 279. P. 1–16. DOI: 10.1134/S0081543812090015
5. Brylawski T. The lattice of integer partitions. *Discrete Math.*, 1973. Vol. 6, No. 3. P. 210–219. DOI: 10.1016/0012-365X(73)90094-0
6. Dong F. M., Koh K. M., Teo K. L. *Chromatic Polynomials and Chromaticity of Graphs*. Hackensack: World Scientific, 2005. 384 p. DOI: 10.1142/5814

7. Farrell E. J. On chromatic coefficients. *Discrete Math.*, 1980. Vol. 29, No. 3. P. 257–264. DOI: 10.1016/0012-365X(80)90154-5
8. Gein P. A. About chromatic uniqueness of complete tripartite graph $K(s, s-1, s-k)$, where $k \geq 1$ and $s-k \geq 2$. *Sib. Èlektron. Mat. Izv.*, 2016. Vol. 13. P. 331–337. DOI: 10.17377/semi.2016.13.027 (in Russian)
9. Gein P. A. About chromatic uniqueness of some complete tripartite graphs. *Sib. Èlektron. Mat. Izv.*, 2017. Vol. 14. P. 1492–1504. DOI: 10.17377/semi.2017.14.129 (in Russian)
10. Gein P. A. On garlands in χ-uniquely colorable graphs. *Sib. Èlektron. Mat. Izv.*, 2019. Vol. 16. P. 1703–1715. DOI: 10.33048/semi.2019.16.120
11. Koh K. M., Teo K. L. The search for chromatically unique graphs. *Graphs Combin.*, 1990. Vol. 6, Suppl. 3. P. 259–285. DOI: 10.1007/BF01787578
12. Koroleva T. A. Chromatic uniqueness of some complete tripartite graphs. I. *Trudy Inst. Mat. i Mekh. UrO RAN*, 2007. Vol. 13, Suppl. 3. P. 65–83. (in Russian)
13. Koroleva T. A. Chromatic uniqueness of some complete tripartite graphs. II. *Izv. Ural. Gos. Univ. Mat. Mekh. Inform.*, 2010. Vol. 74. P. 39–56. (in Russian)
14. Li N. Z., Liu R. Y. The chromaticity of the complete t-partite graph $K(1, P_2 \ldots p_t)$. *J. Xinjiang Univ. Natur. Sci.*, 1990. Vol. 7, No. 3. P. 95–96.
15. Senchonok T. A. Chromatic uniqueness of elements of height 2 in lattices of complete multipartite graphs. *Trudy Inst. Mat. i Mekh. UrO RAN*, 2011. Vol. 17, No. 3. P. 271–281. (in Russian)
16. Zhao H. *Chromaticity and Adjoint Polynomials of Graphs.* Zutphen, Netherlands: Wöhrmann Print Service, 2005. 179 p.
17. Zhao H., Li X., Zhang Sh., Liu R. On the minimum real roots of the σ-polynomials and chromatic uniqueness of graphs. *Discrete Math.*, 2004. Vol. 281, No. 1–3. P. 277–294. DOI: 10.1016/j.disc.2003.06.010

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 66–80
DOI: 10.15826/umj.2021.1.005

SCREENING IN SPACE: RICH AND POOR CONSUMERS IN A LINEAR CITY[1]

Sergey Kokovin[†], **Fedor Vasilev**[††]

HSE University,
3a Kantemirovskaya str., St. Petersburg, 194100 Russia

[†]skokov7@gmail.com, [††]vasilevfyo@gmail.com

Abstract: Unlike standard models of monopolistic screening (second-degree price discrimination), we consider a situation where consumers are heterogeneous not only vertically, in their willingness to pay, but also horizontally, in their tastes or "addresses" a la Hotelling's Linear City. For such a screening game, a novel model is composed. We formulate the game as an optimization program, prove the existence of equilibria, develop a method to calculate equilibria, and characterize their properties. Namely, the solution structure of the resulting menu of contracts can be either a "chain of envy" like in usual screening or a number of disconnected chains. Unlike usual screening, "almost all" consumers get positive informational rent. Importantly, the model can be extended to oligopoly screening.

Keywords: Screening, Price discrimination, Spatial competition, Linear city, Principal-Agent model, Nonconvex optimization.

1. Introduction

Motivation. *In economic practice*, screening or "second-degree price discrimination" is quite usual in many industries. It typically generates a "product line", which is a menu of quantity-price or quality-price "packages". E.g., "packages" can mean various bottles of a soft drink offered to a heterogeneous consumers' population: (300*ml* for \$0.5), (450*ml* for \$0.8), (1000*ml* for \$0.95), and so on. Profit-maximizing product lines in telephony, clothes, cars, everywhere typically demonstrate some price discounts for higher quantity or quality. Why? To explain discounts and to construct product lines, economists exploit knowledge about multiple consumer types, each type being described by its "willingness-to-pay" (monetary valuation function) for higher quantity or quality. Existing types are known, but who belongs to which type is hidden from the seller; consumers self-select based on this asymmetric information — this kind of game is called "screening".

In economic theory, the standard model explaining screening or product lines dates back to Michael Spence [7]. The model is reproduced in many textbooks, monographs [5], and reviews [8]. Typically, this theory focuses on a monopolistic seller and exploits the Spence–Mirrlees "vertical-ordering" assumption: types of consumers are numbered in such a way that a higher number has a higher derivative of its valuation function everywhere. Then, the basic finding is the "Chain-rule theorem" about the list of active constraints in the profit-maximizing menu of packages. This list constitutes a "chain of envy" among consumers: the highest type is almost eager to switch to his/her lower neighbor's package, who in turn is almost eager to switch to his/her lower neighbor's package, and so on, other constraints being redundant. As a result, the solution method is clear

[1]The authors are grateful to Pavel Ilinov, Igor Bykadorov, Mikhail Martyanov, Pavel Molchanov for discussions and help in checking the proofs. The study was financed by the HSE University Basic Research Program.

and the properties of equilibria are definite. Only the highest consumer type gets his/her Pareto-efficient quantity, others are under-served. However, everybody except the lowest type gets some informational rent, i.e., a consumer surplus born by asymmetric information.

Among extensions of this theory, [4] shows that the named properties and the "chain of envy" itself are not guaranteed in situations where the standard Spence-Mirrlees assumption does not hold. Other extensions are also devoted to revealing various solution structures and properties (see [8]).

The present paper considers one more class of non-standard, poorly studied situations. Here the vertical heterogeneity of consumers (i.e., ordered valuations of types) is combined with their horizontal differentiation in some space of tastes or locations. Thus we bridge the theory of screening with another theoretical tradition, horizontal "Linear City". The latter was pioneered by [2] and includes plenty of models (for review, see [9]).

The rationale for combining both "vertical" and "horizontal" theories lies in the realism of the combination. For instance, a typical city includes consumers who are heterogeneous not only in their willingness to pay for quantity/quality but also in their geographical location. When designing his/her product lines, each seller should have in mind not only possible switches of consumers among the packages of his/her product line, but also a possibility to attract more consumers by lower prices of the whole product line. The interaction between these possibly contradicting selling strategies is the main idea here, a novel theoretical question worth studying.

Setting. Our consumers are differentiated in two ways: vertically (e.g., rich and poor) and horizontally. In a certain industry, the horizontal dimension may mean not only geography but also "tastes locations", which include some other characteristic, e.g., size of clothes. Say, T-shirts may be differentiated in qualities (vertical dimension) and sizes (horizontal dimension), suiting various consumer tastes. Still, we stick to expressive geographical interpretation, bearing in mind that tastes interpretation is isomorphic.

In the vertical dimension, we assume a finite number of consumer types. E.g., these types can be "rich", willing to pay more for quality, "middle-class", and "poor". In the horizontal dimension there is a continuum of locations among each type of consumer, uniformly distributed on the real line, the monopolistic seller being located at 0. The consumers bear some transportation costs, paying with their time and effort to go shopping. Therefore, "the farthest customer" is one whose net-of-price willingness-to-pay for the commodity almost equals his/her disutility of walking to the shop. This trade-off generates a negative dependence of the seller's range of service upon its price, interrelated with "envy" among rich and poor.

Results. We propose a novel model for such "Principal-Agents" games and formulate it as the Principal's optimization program under the Participation constraints and Incentive–Compatibility constraints (IC constraints). Such optimization (traditionally) replaces a game among agents. Further, we reduce the Principal's program to a convenient form, show why a solution should exist. Discussing a solution method, we note that our optimization problem need not be a convex one, which brings complications. In principle, one should search among all possible envy structures, which are all possible combinations of constraints. We propose a heuristic directed-search method to determine the set of active constraints in a smaller number of steps than a complete search. It exploits the first-order conditions in such a way as not to miss the global maximum.

The computational issues are the necessary preliminaries, but the economic properties of solutions are the main goal of theoretical studies in screening. For our model, the important finding is that prices tend to be lower than under screening without space. Unlike the standard setting, almost all consumers (except "the farthest customer" of each type) get some informational rent, which is the consumer surplus. Moreover, we characterize the condition on parameters that generate equilibria with Pareto-efficient sizes of packages not only for the highest type but also for some other types. In this case, the usual chain structure of envy among consumer types is broken and the graph of the solution takes the form of a broken chain, some segments being disconnected. Examples illustrate such disconnected equilibria. The highest type in each segment gets the Pareto-efficient size of his/her package, whereas all other agents get distorted sizes. Why should we care about these properties? Because knowledge about Pareto-inefficiency is needed for market regulation, while knowledge about the structure of solutions can help sellers to choose their pricing policies.

Extensions and other approaches. Last but not least motivation: the present setting is one of the possible ways to develop screening in oligopoly, which remains a difficult goal for theorists to reach. In this field, the well-known paper by Rothschild and Stiglitz [6] is probably a good description of a competitive market of insurance, but it does not fit well the commodity markets because it predicts zero profit for sellers, which is not realistic. Our spatial approach appears as a good alternative because space softens competition and allows for realistic positive profits. Another approach to screening in oligopoly was recently developed in an important paper by Chade and Swinkels [1]. Their firms are heterogeneous; as a result, higher types of firms serve higher consumer types (sorting). Payoffs in such a game are not quasi-concave, but equilibria do exist. Certain firm types distort their allocations downwards; the welfare effects of private information differ from those under monopoly. The approach by Chade and Swinkels is an alternative (non-spatial) way to describe oligopoly screening with positive profit, which complements our approach.

The next section introduces the model. Subsequent sections provide examples, show that equilibria exist, provide a method to find them, and discuss equilibria properties.

2. Screening in a "Linear City": model and equilibria existence

"Linear City" in theory is typically a continuously inhabited interval, or, like here, the real line $(-\infty, \infty)$, where location 0 is the "city center". For simplicity, we restrict our attention to *uniform* distribution of each consumer type. Our single seller is located in the center, at 0. So, by symmetry, it is sufficient to formally represent only one side of this city $[0, \infty)$, the other side $(-\infty, 0]$ just mirrors the first one.[2]

Operating at 0, our monopolistic seller (a shop) serves the interval $[0, \infty)$ uniformly populated by each of n consumer types. Each individual is characterized by his/her type $i \in \{1, 2, ..., n\}$ and by location $\xi_i \in [0, \infty)$, which is his/her distance from the seller. The seller offers at 0 some commodity or service to all consumers, constructing a menu $\{(q_1, T_1), ..., (q_n, T_n)\}$, where each package includes quality/quantity/size q_i of the commodity and tariff or price T_i. (For instance, one may think of a shop selling a soft drink. Then, the size q_1 is destined to the "least thirsty" consumers, $q_2 > q_1$ should serve "moderate thirsty" ones, while $q_3 > q_2$ should serve "very thirsty" consumers, or very rich ones, eager to pay more.) The "transportation cost" $\tau\xi_i$ is proportional to distance (with $\tau > 0$). It is interpreted as the customers' time/money, spent traveling for shopping at point 0.[3]

Each point is inhabited by some *mass* $m_1 > 0$ of consumers type #1, by some mass $m_2 > 0$ of consumers type #2, mass $m_3 > 0$ of consumers type #3, and so on. (If one interpret m_i

[2]Choosing an infinite city instead of $[-1, 1]$ means that we want to model sufficiently *high transportation costs*, to make the city edges *not served* by the seller. We are going to investigate also the case of a city completely covered by service, and the most interesting topic — two oligopolists, competing for the whole Linear City.

[3]As we had said, another typical interpretation of Linear City $[0, \infty)$ is some space of characteristics, e.g., sizes of clothes. In this case, the "transportation cost" means *disutility* from inappropriate size, expressed in money.

as a *probability* of this type, then normalization $m_1 + m_2 + ... + m_n = 1$ should be added, but normalization plays no role in our solution.) The consumers are characterized by their *monetary valuation functions,* i.e., willingness to pay for quantity/quality, denoted by $\tilde{v}_1[q]$, $\tilde{v}_2[q]$, $\tilde{v}_3[q]$.

On the other side of the counter, the "Principal" in this game, the monopolistic seller, faces a *marginal cost* $c \geq 0$, which is his/her production cost per unit of quantity/quality. However, we shall subtract costs from valuations/tariffs and further deal only with *net-of-cost valuations* v *and net-of-cost tariffs* t[4]:

$$v_1[q] \equiv \tilde{v}_1[q] - cq, \quad \ldots, \quad v_n[q] \equiv \tilde{v}_n[q] - cq[q],$$
$$t_1 \equiv T_1 - cq_1, \quad \ldots, \quad t_n \equiv T_n - cq_n.$$

Respectively, the subsequent analysis looks as if costs were zero ($c = 0$), but the non-zero case is also included in consideration because the summand cq_i enters both sides of important constraints and cancels out.

Assumption 1 (*Boundedness+*). *Each net valuation* $v_i : R_+ \to R_+$ $(i = 1, 2, \ldots, n)$ *has a finite argmaximum*:

$$\exists\ q_i^o \equiv \arg\max_z v_i[z] > 0.$$

Each net valuation function v_i *is strictly concave, twice continuously differentiable, and it is strictly increasing on* $[0,\ q_i^o)$, *i.e., below the argmaximum.*

Assumption 2 (*Ordering+*). *The family of valuations* $v_1[\cdot], v_2[\cdot], \ldots, v_n[\cdot]$ *satisfies the Spence–Mirrlees ordering condition*:

$$v_1'[q] < v_2'[q] < v_3'[q] \quad \forall q, \quad v_i[0] = 0 \quad \forall i = 1, 2, \ldots, n.$$

As a result, graphs of all net valuations $v_i[q]$ do cross at the origin and never cross again, that is why such assumption is often called the "single-crossing condition".

Example 1. An example of valuations' family, used below for demonstrations, is a family of *affine transforms* of some common function $\nu[\cdot]$:

$$v_1[q] = a_1 q + \nu[q], \quad v_2[q] = a_2 q + \nu[q], \quad v_3[q] = a_3 q + \nu[q],$$

with some parameters $0 < a_1 < a_2 < a_3$. E.g., it can be a family of parabolas like

$$v_1[q] = 2q - 0.5q^2, \quad v_2[q] = 3q - 0.5q^2, \quad v_3[q] = 4q - 0.5q^2,$$

see similar examples below.

Traditionally for the screening theory, in such games, consumers play the role of informed "Agents", or "followers". The seller, uninformed about their types, is a "Principal", or "leader": he/she plays first, they second. He/she needs to construct a *menu of* "packages", being unable to discriminate among their types. Each package $(q_i, t_i) \geq 0$ includes quality q_i and tariff t_i, called also "price". One can show that there is no need to construct more packages than n agent types in the market. Non-participation is perceived as one more package $(0, 0)$. So, the menu will consist of $(0, 0)$ and n non-trivial elements

$$(q, t) = ((q_1, t_1), (q_2, t_2), \ldots, (q_n, t_n)).$$

[4]Hereinafter, we always use brackets like $f[\cdot]$ to denote arguments of functions, using parentheses $(\cdot)$ for grouping.

If some packages coincide, then the consumers actually have less than n different options, but formally we shall discuss exactly n packages designed.

After the menu is set, each agent comes and "self-selects", i.e., buys a single package (q_i, t_i) from the menu in take-it-or-leave-it fashion. He/she can choose the *zero outside option* (not buy anything), then (0,0) brings him/her zero "reservation utility" $u_{i0} = 0$. Whenever any agent (i, ξ) chooses a package (q_i, t_i), his/her payoff, or gross utility u_i will include his/her valuation $v_i[q_i]$ for the chosen quality q_i minus the tariff t_i, minus his/her personal transportation cost, as follows:

$$u_i[q_i, t_i, \xi] = v_i[q_i] - t_i - \tau\xi \geq u_{i0} = 0.$$

Here, the parameter $\tau > 0$ is the "distance cost" coefficient for any consumer located at ξ, i.e., ξ-far from the seller located at 0. In other words, the farther is the consumer from the shop, the more he/she spends on shopping. For any type i, the endogenous *range of service* x_i is defined as the location of the *farthest consumer* among this type who comes to buy anything:

$$x_i : \left(q_i[\xi_i] > 0 \quad \forall \xi_i \leq x_i\right), \quad \left(q_i[\xi_i] = 0 \quad \forall \xi_i > x_i\right).$$

In other words, all ξ_i located closer than x_i to 0 do buy from the seller, more distant consumers do not (the ranges of service x_i may be different among types i).

We have normalized the marginal cost c to zero (without loss of generality). So, the Principal's elementary payoff from a single purchase by consumer (i, ξ) is

$$\pi_i[q_i, t_i] \equiv m_i * (t_i - cq_i) = m_i * t_i.$$

Taking into account the endogenous range of service $x_i = x_i\,[q_i, t_i] \geq 0$, we are going to maximize the total Principal's expected profit, which is the integral over all consumers served (the weighted sum of individual net tariffs)

$$\Pi = \Pi[q, t, x] \equiv x_1 \cdot m_1 \cdot t_1 + ... + x_n \cdot m_n \cdot t_n \to \max_{\{(x_i, q_i, t_i)_{i \leq n} \geq 0\}}. \tag{2.1}$$

One may be surprised that the Principal is expected to design not only the packages but also the range of service x_i. Let us explain: traditionally for such theory, the consumers goals and behavior are expressed through inequalities. Namely, profit is maximized w.r.t. all variables simultaneously, including those chosen by consumers, under two groups of agents "rationality constraints". These are almost-standard Participation constraints (2.2) and Incentive-compatibility constraints (2.3):

$$[q_i] - t_i \geq \tau x_i \quad \forall i, \tag{2.2}$$
$$v_i[q_i] - t_i - \tau x_i \geq v_i[q_j] - t_j - \tau x_i \quad \forall i, j. \tag{2.3}$$

Here constraint (2.2) means that the consumer's surplus from the purchase exceeds his/her transport cost. It includes the novelty of our model: without the spatial dimension, the right-hand side of the participation constraints would be just zero. We also have in mind participation constraints

$$v_i[q_i] - t_i \geq \tau\xi_i \quad \forall \xi_i < x_i$$

for all close-to-producer agents (i, ξ), but they are weaker than such constraint (2.2) for the farthest consumer x_i, and therefore dropped.

Any Incentive-compatibility constraint (2.3) means that a consumer i is not "envying" any other (jth) package, i.e., he/she has no incentive to take package j instead of one designed for him/her. The transport cost τx_i enters both sides of the Incentive Compatibility inequality, so, it can be dropped.

Now we show that the usual Chain–Rule applies here, i.e., that many Incentive-compatibility constraints can be dropped or replaced.

Lemma 1 (Chain-Rule). *Solving the maximization problem* (2.1)–(2.3) *is equivalent to maximizing the same objective function under the following constraints*:

$$v_1[q_1] - t_1 \geq \tau x_1, \quad \dots, \quad v_n[q_n] - t_n \geq \tau x_n, \tag{2.4}$$
$$v_2[q_2] - t_2 \geq v_2[q_1] - t_1, \tag{2.5}$$
$$v_3[q_3] - t_3 \geq v_3[q_2] - t_2, \tag{2.6}$$
$$\dots$$
$$v_n[q_n] - t_n \geq v_n[q_{n-1}] - t_{n-1}, \tag{2.7}$$
$$q_n \geq ... \geq q_2 \geq q_1 \geq 0, \tag{2.8}$$

where some Incentive-compatibility constraints are replaced by the ordering constraints (2.8).

P r o o f. We observe that our initial optimization problem (2.1)–(2.3) differs from the classical one only in its participation constraints. This allows us to repeat the classical proof under the Spence–Mirrlees condition (see [3]), and claim that a constraint "no-envy from any i to his/her lower neighbor $i-1$" *implies* "no-envy" from i to anybody else:

$$(v_i[q_i] - t_i \geq v_i[q_{i-1}] - t_{i-1}) \Rightarrow v_i[q_i] - t_i \geq v_i[q_j] - t_j \quad \forall j.$$

First, we combine constraint of "no-envy" from any i to any lower $j < i$ with its inverse: "no-envy" from j to the higher type i:

$$v_i[q_i] - v_i[q_j] \geq t_i - t_j \geq v_j[q_i] - v_j[q_j] \tag{2.9}$$

and compare this with the Spence–Mirrlees condition expressed in finite differences:

$$v_i[q_i] - v_i[\tilde{q}] \geq v_j[q_i] - v_j[\tilde{q}] \;\; \forall j < i, \forall\, (q_i, \tilde{q}) \,|\, q_i \geq \tilde{q}.$$

We conclude that all incentive-compatible packages must satisfy the q-ordering constraint (2.8), i.e., a higher type must take a (weakly) bigger package. Thereby, adding this ordering constraint to the constraints system (2.2)–(2.3) does not influence optimization. Since our objective function is increasing in t_i and our constraints take the form $v_i[q_i] - t_i \geq ...$, we realize that it is sufficient to consider only intervals $q_i \leq q_i^o$ below the argmaximum, where our valuations are increasing. Then, it is easy to check that bigger packages imply weakly higher tariffs for higher types: $t_1 \leq t_2 \leq ... \leq t_n$ at any solution. Hence, we can ignore in optimization each "no-envy" constraint from $j < i$ to higher i. Indeed, it is the right inequality in (2.9), whereas a profitable increase in both tariffs t_i, t_j can make only the left equality binding, not the right one.

Second, similarly using q-ordering and the Spence–Mirrlees condition, we check that "no-envy from any i to his/her lower neighbor $i-1$" implies also "no-envy" from i to any lower type $j < i-1$. Thereby, under (2.8) all non-neighboring incentive constraints are excessive, can be dropped without changing our optimization. □

Thus, we have introduced a new model of screening, and represented a related Principal–Agent game as the Principal's optimization program (2.1), (2.4)–(2.8); all equilibria of our game (if any) are some profit-maximizing solutions.

3. Reduction of variables and existence of solutions

To reduce variables, one can look at the objective function (2.1) increasing in t_i, x_i, and conclude that, for each type i, the farthest-customer's participation constraint ***must be active*** at any solution, i.e., become an equality (in the opposite case we could increase profit by increasing variables t_i, x_i).

So, we can use these active constraints to express the variables t_i as

$$t_1 = v_1[q_1] - \tau x_1, \quad t_2 = v_2[q_2] - \tau x_2, \quad \dots, \quad t_n = v_n[q_n] - \tau x_n.$$

Plugging these t_i into the IC constraints, we obtain their left-hand sides equal to the ranges of service x_i:

$$v_2[q_2] - t_2 = v_2[q_2] - (v_2[q_2] - \tau x_2) \equiv \tau x_2, \quad \dots, \quad v_n[q_n] - (v_n[q_n] - \tau x_n) \equiv \tau x_n.$$

Now we plug these expressions into the profit Π and into the IC constraints, thereby excluding the variables t_i and participation constraints. Thus we come to the reduced maximization problem to be solved:

$$\Pi \equiv x_1 \cdot m_1 \cdot (v_1[q_1] - \tau x_1) + ... + x_n \cdot m_n \cdot (v_n[q_n] - \tau x_n) \to \max_{(x_i, q_i)_{i \le n}} \text{ s.t.} \tag{3.1}$$

$$\begin{aligned} \tau x_2 &\geq v_2[q_1] - v_1[q_1] + \tau x_1, \\ \tau x_3 &\geq v_3[q_2] - v_2[q_2] + \tau x_2, \\ &\dots \\ \tau x_n &\geq v_n[q_{n-1}] - v_{n-1}[q_{n-1}] + \tau x_{n-1}, \\ q_n &\geq q_{n-1} \geq ... \geq q_1 \geq 0. \end{aligned} \tag{3.2}$$

Possible solution "structures" and ideas of solving. It is common in constrained optimization to find a solution through exploring many possible combinations of constraints — inequalities, when finding out which of them will become *active* (equalities) at the true global maximum. In convex optimization, e.g., linear programming, well-known are algorithms of *directed* search among these combinations. An efficient directed search reduces the number of combinations explored, keeping a warranty of the true optimum. We are going to construct a sort of such search here. We shall denote by

$$A = \{IC_{ij}, IC_{jk}, ..., O_i, ...\}$$

any possible "solutions structure", i.e., the list (combination) of names of constraints that we assume are active at the current step. Hereinafter, IC_{ij} denotes the Incentive Compatibility constraint like $\tau x_i \geq v_i[q_j] - v_j[q_j] + \tau x_j$, and an ordering constraint like $q_i \geq q_{i-1}$ is denoted by O_i. Under any A-hypothesis, we call a related solution an "A-conditional optimum". After trying all A, we compare all such conditional optima to select a true optimum.

Unfortunately, in general, our optimization program need not be a convex one! Indeed, one can note that our objective function includes the summands $x_1 \cdot v_1[q_1]$ where both multipliers are increasing, this form precludes concavity of this function. Moreover, our constraints include the difference $v_{i+1}[q_i] - v_i[q_i]$ (of concave functions) that need not be convex or concave without an additional assumption. Generally, our domain for variables is not necessarily convex. However, our specific problem often allows for some simplifications. We start discussing them with possible empty set of active constraints $A = \emptyset$.

Disconnected kind of solutions. To introduce additional notions and notations before our existence theorem, we now show some specific, "disconnected" type of solutions (equilibria), that may occur under some specific valuations v_i.

Let us suppose that all incentive constraints (IC) and all ordering constraints $q_i \geq q_{i-1}$ are *inactive,* play no role in the solution. Then, optimization in q alone would give us the so-called

"Pareto-optimal" package sizes q_i^o, because they maximize $v_i[q_i]$ *per se*:

$$q_3^o \equiv \arg\max_{z\geq 0} v_3[z] \quad \geq q_2^o \equiv \arg\max_{z\geq 0} v_2[z] \geq q_1^o \equiv \arg\max_{z\geq 0} v_1[z] > 0, \tag{3.3}$$

$$t_3^o \equiv \max_{z\geq 0} v_3[z]/2 \quad \geq t_2^o \equiv \max_{z\geq 0} v_2[z]/2 \geq t_1^o \equiv \max_{z\geq 0} v_1[z]/2 > 0. \tag{3.4}$$

Using these known values q, t, now we can optimize in x_i each summand $x_i \cdot m_i \cdot (v_i[q_i^o] - \tau x_i)$ and obtain "Pareto-optimal" ranges x_i^o of service

$$x_1^o \equiv \frac{v_1[q_1^o]}{2\tau}, \quad x_2^o \equiv \frac{v_2[q_2^o]}{2\tau}, \quad \dots \quad x_n^o \equiv \frac{v_n[q_n^o]}{2\tau}. \tag{3.5}$$

Such a solution means that, out of common benefit $v_i[q_i^o]$ from their contract, the closest-to-the-seller consumer gets one half (as "consumer surplus"), and the seller gets the other half as his/her profit. More distant consumers get less.

To find what kind of valuations may generate such disconnected equilibria, we plug these expressions q_1^o, t_1^o into the incentive constraints. Thus we get a necessary condition (3.6) on such valuations:

$$\frac{v_{i+1}[q_{i+1}^o]}{2} - \frac{v_i[q_i^o]}{2} \geq v_{i+1}[q_i^o] - v_i[q_i^o] \Rightarrow$$
$$v_{i+1}[q_{i+1}^o] \geq 2v_{i+1}[q_i^o] - v_i[q_i^o] \tag{3.6}$$

for all i. Is the inequality plausible, is a *disconnected solution possible* under any valuations v_i? The following example confirms this.

Example 2. (Separated types #1, #2, #3.) The following example with three quadratic valuations $v_i[q_i] = a_i * q_i - b_i * q_i^2$ shows a disconnected structure:

$$a_1 = 2; \quad a_2 = 2.2; \quad a_3 = 2.3; \quad b_1 = 5; \quad b_2 = 2; \quad b_3 = 0.8.$$

The masses of types are $m1 = m2 = m3 = 1$, and the costs are $c = 0$; $\tau = 1$.

These data and direct calculations yield the following profit-maximizing Pareto-optimal sizes/tariffs:

$$q_1^* = q_1^o \equiv \arg\max_q v_2[q] = 0.2, \quad t_1 = 0.1,$$
$$q_2^* = q_2^o \equiv \arg\max_q v_2[q] = 0.55, \quad t2 = 0.3025,$$
$$q_3^* = q_3^o \equiv \arg\max_q v_3[q] = 1.4375, \quad t_3 = 0.826563.$$

Fig. 1 exhibits our Example 2 with a disconnected equilibrium. It gives also some geometry intuitions for this kind of solutions and for our optimization problem *per se*.

Thick red, green, and blue dots are the consumers' equilibrium packages: the quantity q_i lies on the horizontal axis, the tariff t_i lies on the vertical one. The valuations $v_i[\cdot]$ of the first, second, and third consumers are the solid curves painted red, green, and blue, respectively. Each dashed curve shows the equilibrium level of the valuation function for one consumer among this type, namely, for one closest to the seller. The lower is the dashed curve, the better for the consumer because its difference in height with related solid curve demonstrates the consumer's surplus (payoff).

Small red, green, and blue squares demonstrate the equilibrium packages which would occur under standard, space-less screening. Comparing standard and new outcomes, we observe that

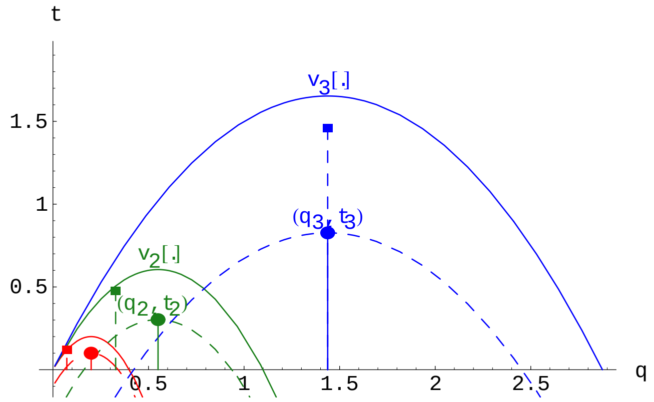


Figure 1. Valuations and solutions.

adding space to screening can diminish distortion. Namely, space increases quantities q and also brings benefits for the close-to-seller consumers by diminishing tariffs t.

Example 3. (Connected types #1+#2, separated #3.) This example is almost the same as Example 2, only $b_2 = 3$. The quadratic valuations are

$$v_i[q_i] = a_i * q_i - b_i * q_i^2$$

with

$$a_1 = 2; \quad a_2 = 2.2; \quad a_3 = 2.3; \quad b_1 = 5; \quad b_2 = 3; \quad b_3 = 0.8,$$

the same masses $m1 = m2 = m3 = 1$ and costs $c = 0$; $\tau = 1$. The solution shows a partially disconnected structure:

$$q_1^* = 0.19 < \arg\max_q v_2[q], \quad t_1 = 0.1104336,$$

$$q_2^* = q_2^o \equiv \arg\max_q v_2[q] = 0.366667, \quad t_2 = 0.19712,$$

$$q_3^* = q_3^o \equiv \arg\max_q v_3[q] = 1.4375, \quad t_3 = 0.826563.$$

Example 4. (Connected types #1+#2+#3.) This example is almost the same as Example 3, only $b_3 = 2$. The quadratic valuations are

$$v_i[q_i] = a_i * q_i - b_i * q_i^2$$

with

$$a_1 = 2; \quad a_2 = 2.2; \quad a_3 = 2.3; \quad b_1 = 5; \quad b_2 = 3; \quad b_3 = 2,$$

the same masses $m1 = m2 = m3 = 1$ and costs $c = 0$; $\tau = 1$. It shows a completely connected solution structure:

$$q_1^* = 0.182261 < \arg\max_q v_2[q], \quad t_1 = 0.10786,$$

$$q_2^* = 0.34472 < \arg\max_q v_2[q], \quad t_2 = 0.208432,$$

$$q_3^* = q_3^o \equiv \arg\max_q v_3[q] = 0.575, \quad t_3 = 0.31449.$$

Economic intuitions. Economically, why the "chain of envy", quite usual in standard screening, can be broken in spatial screening? Why disconnected profit-maximizing solutions may happen? Well, we know that in usual screening there are no reasons for the principal to leave any incentive compatibility constraint, or the 1st participation constraint, inactive. Such a plan would bring waste of profit. Respectively, the 1st type in usual screening gets zero informational rent, nothing beyond the reservation utility.

By contrast, in our spatial screening model, all closer-than-the-farthest customers get equilibrium payoffs *higher* than their reservation ones; i.e., the tariff for them is lower than it could be. This slack is not wasted, from the view of the profit-maximizing Principal, it is a sacrifice for extending his/her service range, his/her coverage of consumers. This objective, *constructing a utility slack to attract more consumers*, can make one or more incentive constraints inactive. The realism of such a trade-off in many markets — is the main reason for building our new model of screening.

Our discussion of examples and reasons for non-active constraints ends up with the following conclusion. Generally, the list A of active constraints may be empty, or include all IC constraints ($A = \{IC_{n,n-1}, IC_{n,n-1}, ..., IC_{21}\}$), or may consist of various combinations of active constraints.

Equilibria existence. Returning to the general case with unknown A, we should ask: are there always solutions $\{(x_i, q_i)_{i \le n}\}$ to our maximization program (3.1)–(3.2)? Our domain is not empty. Indeed, to show a *sample admissible plan* we can take all partial Pareto-optimal sizes q_i^o. These "separate maxima" exist under our assumptions, they satisfy our ordering.[5] We can supplement these q with sufficiently small x like $x_1 = x_2 = x_3 = \varepsilon < \min_j v_j[q_j^o]/\tau$. So, a positive admissible plan $(q, x) > 0$ exists. Moreover, it brings a strictly positive profit $\Pi > 0$.

Now studying our objective function, we observe that the highest-type quality q_n enters Π only once, so, this variable must take its Pareto-optimal value $q_n = q_n^o$ at any solution. As to the other arguments q_i of the objective function, we conclude that we can restrict them as $q_i \le q_i^o$ without sacrificing our objective function (since, for higher values of q, our profit Π becomes smaller). Similarly, without sacrificing our objectives, variables x can be restricted as $(v_i[q_i] - \tau x_i) \ge 0$, i.e., $x_i \le v_i[q_i]/\tau \le v_i[q_i^o]/\tau$.

Therefore, without loss of optima, we can squeeze our initial domain (positive orthant), and maximize now our function $\Pi(q, x)$ on a restricted compact domain K constructed as

$$K \equiv \big\{(q, x) | \, 0 \le q_i \le q_i^o + 1, \quad 0 \le x_i \le x_{max} \equiv \max_j v_j[q_j^o]/\tau + 1 \quad \forall i = 1, 2, 3\big\}.$$

Our objective function $\Pi(q, x)$ is continuous. Therefore, by Weierstrass's extreme value theorem, we have established the following statement.

Proposition 1. *Under our assumptions, there exists a solution to our maximization problem* (3.1)–(3.2).

Moreover, one can conclude from our discussion that the maximum lies strictly below the artificial upper bounds $q_i^o + 1$, x_{max} (being an inner solution) and brings a positive profit Π.

How can we practically find a solution under any specific valuation functions v_i?

Of course, one can exploit any iterative or exact numerical method. Since we deal here with differentiable functions, it is possible to use exact finite methods, exploiting the first-order conditions, even without being sure in convex optimization. Indeed, after finding all stationary points and border solutions, one can compare (a finite number of) related local maxima, to choose the global one.

[5] We also can take sufficiently small values of $x_1 = x_2 = x_2 = \varepsilon < v_1[q_1^o]/\tau < v_2[q_2^o]/\tau < v_3[q_3^o]/\tau > 0$, and assure that the objective function *can be positive* under some values of the optimizers.

To implement this idea, solving first-order equations for *all* possible A-structures should be sufficient to find the true optimum, but this way can be computationally tedious. We shall suggest a heuristic method that explores *some* A-structures. To construct each step of this method, the next section suggests a convenient way of using First-Order Conditions (FOC) for exploring any hypothesis A about active constraints.

4. Using FOC for any hypothetical list A of active constraints

This section explains how to further reduce our variables and use First-Order Conditions to find any A-conditional maximum under some hypothesis A. We explain it by an example.

Suppose that we have $n = 4$ consumer types and assume that family A of active IC constraints (A explored on some step of our general search algorithm) connects three adjacent agent types $\{\#1, \#2, \#3\}$ as $A = \{IC_{21}, IC_{32}\}$, whereas type #4 is separated and ordering constraints are not binding. We find related A-conditional maximum as follows.

As we have ensured, the highest variable q_3^* among the chain $\{\#1, \#2, \#3\}$ must take its v_3-maximizing Pareto-value q_3^o (here and further accent * denotes solutions):

$$q_3^* = q_3^o \equiv \arg\max v_3[q_3].$$

Similarly, the isolated variable q_4 also takes its Pareto-optimal value $q_4^* = q_4^o$. (By contrast, lower variables q_i in the chain need not become Pareto-optimal because the incentive constraints may be active.)

Whenever any IC_{ji} is active, we can define the difference function $V_{ji}[q_i] \equiv v_j[q_i] - v_i[q_i]$. E.g., for two IC included in $A = \{IC_{21}, IC_{32}\}$, these difference functions are

$$V_{32}[q_2] \equiv v_3[q_2] - v_2[q_2], \quad V_{21}[q_2] \equiv v_2[q_1] - v_1[q_1].$$

(**Special linear case**. Such difference function $V_{32}[\cdot]$ can appear *linear* in the particular case when the valuations family $v_i[\cdot]$ is built from a common function $u[\cdot]$ as its linear modification $v_i[x] = a_i x - u[x]$ with $a_1 < a_2 < ....$ In this case, $V_{32}[q_2] \equiv a_3 q_2 - a_2 q_2 = (a_3 - a_2) q_2$, $V_{21}[q_1] \equiv a_2 q_1 - a_1 q_1 = (a_2 - a_1) q_2$).

We can invert any difference function V_{ij} because it is *increasing,* by Spence–Mirrlees assumption. We denote the inverse $\Lambda_{ij}[\cdot] \equiv V_{ij}^{-1}[\cdot]$. Further, to reformulate active constraints — equations $\tau x_3 = v_3[q_2] - v_2[q_2] + \tau x_2$ through these functions Λ_{ij}, we express the volumes q_i through the differences in service ranges:

$$\begin{aligned} q_2 &= \Lambda_{32}\left[\tau x_3 - \tau x_2\right] \equiv V_{32}^{-1}\left[\tau x_3 - \tau x_2\right], \\ q_1 &= \Lambda_{21}\left[\tau x_2 - \tau x_1\right] \equiv V_{21}^{-1}\left[\tau x_2 - \tau x_1\right]. \end{aligned}$$

Using this transform to simplify our optimization, we can get rid of all variables except the service ranges x_i:

$$\begin{aligned} \tau\Pi = \tau x_1 \cdot m_1 \cdot (v_1[\Lambda_{21}\left[\tau x_2 - \tau x_1\right]] - \tau x_1) + \tau x_2 \cdot m_2 \cdot (v_2[\Lambda_{32}\left[\tau x_3 - \tau x_2\right]] - \tau x_2) + \\ \tau x_3 \cdot m_3 \cdot \left(v_3[q_3^{opt}] - \tau x_3\right) \to \max_{x=(x_1,x_2,x_3)\geq 0}, \end{aligned} \tag{4.1}$$

and deal with unconstrained optimization. We check after finding the unconstrained maxima whether the ordering conditions and out-of-A IC constraints are satisfied. In the opposite case (violated outside constraints), we *reject* the hypothesis A and explore another one.

In formulation (4.1), we have multiplied our objective function by τ to prepare subsequent usage of auxiliary variables $y_i \equiv \tau x_i$, δ_{ij}, to simplify the analysis. This trick explains also the following remark.

Remark 1. The parameter τ only multiplies our payoff but does not influence the main variables — maximizers (x, q, t).

Solution for a connected component $\{\#i, \#i+1, \#i+2\}$ under $A = \{IC_{21}, IC_{32}\}$. To find the solution components (x_1, x_2, x_3), it is sufficient to differentiate the partial objective function (4.1) by (x_1, x_2, x_3), explore these first-order conditions, and compare all resulting stationary points (the points can be multiple if the function is non-concave, but their number is finite), to choose the true maximum. Afterwards we derive the remaining variables (q, t) from these (x^*). This gives the part $(q_i^*, t_i^*, x_i^*)_{i\leq 3}$ of the needed solution. Turning to the remaining type #4, the part $(q_i^*, t_i^*, x_i^*)_{i\leq 3}$ is supplemented with the Pareto-optimal values $(q_4^*, t_4^*, x_4^*) = (q_4^o, t_4^o, x_4^o)$ found for any isolated type from equations (3.3)–(3.5). Now we should check if the ordering conditions and the unused IC constraints (IC_{43}) are really satisfied, inactive. If it is wrong, this hypothesis $A = \{IC_{21}, IC_{32}\}$ is rejected, otherwise, it can be compared with other hypotheses.

To simplify using the first-order conditions, the objective function can be expressed in new auxiliary variables $\delta_{ji} \equiv y_j - y_i$ as:

$$\begin{aligned}\tau\Pi = (y_3 - \delta_{32} - \delta_{21}) \cdot m_1 \cdot v_1[\Lambda_{21}[\delta_{21}]] - m_1 \cdot (y_3 - \delta_{32} - \delta_{21})^2 +\\ (y_3 - \delta_{32}) \cdot m_2 \cdot v_2[\Lambda_{32}[\delta_{32}]] - m_2 \cdot (y_3 - \delta_{32})^2 +\\ (y_3) \cdot m_3 \cdot v_3[q_3^{opt}] - m_3 \cdot y_3^2 \to \max_{y=(\delta_{21},\delta_{32},y_3)\geq 0}.\end{aligned} \tag{4.2}$$

If we treat the variable y_3 parametrically, it is easy to observe that the *concavity* in the remaining variables δ_{ij} of any summand is guaranteed when every function $v_i[\Lambda_{i+1,i}[z]]$ is concave for all z. Concavity may help in practical optimization as well as the following technical lemma.

Lemma 2 (Existence of solutions for a component). Given a family A of active IC constraints and its connected component $\{\#i, \#i+1, \#i+2\}$, the first-order conditions for related function formulated as (4.1) must give at least one solution that is a global maximum of this function.

P r o o f. Though Proposition 1 states that maxima do exist when the problem is expressed in terms of variables (q, x), the above lemma is more specific. Looking at the objective function (4.2) we note that values of $v_i[\Lambda_{i+1,i}[z]]$ are bounded from above by the maximal value $\max_{z\geq 0} v_i[z]$, whereas other terms are quadratic with minus and take arbitrarily low (negative) values when any variable approaches infinity. So, in spite of the unknown concavity of this objective function, we are sure that all its local maxima must be inner ones, not go to infinity. □

As to the *uniqueness* of a solution for any connected component, probably, it can be proved using the assumption of "special linear case", i.e., $\Lambda_{21}'' = 0$, but this question remains unclear.

Now we explain how to use our A-conditional solutions and connected components to sequentially search among A-conditional maxima for finding a true maximum. Related heuristic computational procedure hopefully economizes calculations. The next section also describes all possible solution structures.

5. Non-active IC constraints and method of search among broken chains

We have shown how to find conditional optima for each connected component belonging to any hypothetical family A of active constraints. Now let us show how to go step by step from one hypothetical family A to another, revealing which constraints *should be active* at the solution, i.e., building a sequence of families that approaches the optimal family A^*.

To begin with, we can try using Pareto-optimal sizes $q_i^* = q_i^o \equiv \arg\max_q v_i[q]$ with related profit-maximizing tariffs $t_i^* = t_i^o \equiv v_i[q_i^o]$ and ranges $x_i^* = x_i^o = v_i[q_i^o]/2$ as in (3.3)–(3.5). It can happen that no IC or ordering constraints are violated. Then we can declare that it is a global optimum with a *disconnected solution structure*: $A = \emptyset$, like our Example 2 (because adding any constraint to A cannot enhance the objective function).

More typically, some constraints are violated at $A = \emptyset$, then some packages should become connected, and finding the optimal solution structure A^* becomes more difficult. These considerations give us intuition for the following algorithm for finding A^*, through checking active IC constraint.

The idea of general optimization algorithm.

Let us denote by $ICO \equiv \{IC_{21}, IC_{32}, ..., IC_{n-1,n}, O_1, ..., O_n\}$ the list of all possible Incentive-compatibility and Ordering constraints in our reduced program (3.1). Now we describe an algorithm of directed search among multiple possible combinations, various families $A \subset ICO$ of *active* constraints. Each family A may generate its own *A-conditional-optimal* solution (q^A, x^A), i.e., the solution under these constraints only. If it happens that this solution (q^A, x^A) does not violate other, non-included constraints $ICO \setminus A$, then we have reached an *admissible A-conditionally-optimal* solution (local maximum). Otherwise, we reject the family A as a possible generator of solutions. After we explore ALL *admissible A-conditionally-optimal* solutions for all possible $A \subset ICO$ (exploring finite number of combinations), we can compare their profits and choose the best local maximum, being sure that it is a global maximum.

Computationally, it appears a tedious, long search. However, luckily, our specific optimization problem allows for shorter, sequential, directed search among all possible families A, starting with an empty set $A = \emptyset$ and then adding the active constraints one-by-one, going from lower to higher consumer types, as follows. We provide reasons why during this search we cannot miss the optimal system A^* of active constraints.

Heuristic algorithm of sequential search among possible A-structures.

Let us denote packages by $w_i = (q_i, t_i)$.

1. We start from the lowest type #1. To assign his/her package w_1, we first assume that his/her upper IC_{21} is inactive (separated #1) and therefore assign the related Pareto-optimal values

$$q_1 = q_1^0, \quad t_1 = v_1[q_1^0]/2, \quad x_1 = \frac{v_1[q_1^0]}{2\tau}.$$

2. Similarly, we find the second Pareto-optimal package $w_2 = (q_2, t_2)$, assuming that both IC_{21} and IC_{32} are inactive:

$$q_2 = q_2^0, \quad t_2 = v_2[q_2^0]/2, \quad x_2 = \frac{v_2[q_2^0]}{2\tau}.$$

3. Now we check whether these two packages w_1 and w_2 violate the constraint IC_{21}.

   If IC_{21} is violated, then type 1 and type 2 are "connected", i.e., their packages w_1, w_2 should be optimized together, within one problem in the way explained in Section 4. In this case, we solve program (4.2) that includes these two types: $A = \{IC_{21}\}$. We already know that among these agents the highest size q_2 necessarily becomes Pareto-optimal: $q_2^* = q_2^o$, but the tariff t_2 will differ from Pareto-optimal $v_2[q_2^0]/2$. Anyway, we get some partial plan — an admissible couple (w_1, w_2) .

4. Now we find the third-type optimal package under the assumptions that IC_{32} and IC_{43} are inactive and #3 is separated. Thereby, w_3 should be Pareto-optimal:

$$q_3 = q_3^0, \quad t_3 = v_3[q_3^0], \quad x_3 = v_3[q_3^0]/2.$$

Now we check (violated or not) the constraint IC_{32}, using the packages w_3 and w_2 already found previously. If IC_{32} is violated, then type 2 and type 3 become connected. They become parts of a unified optimization problem with $A = \{IC_{32}, IC_{21}\}$ in the case if 1 and 2 were connected. At this stage, we apply program (4.2) and a related known method to the family $A = \{IC_{32}, IC_{21}\}$ and find three connected packages (w_1, w_2, w_3). Thereby, our previously found (w_1, w_2) will change.

In the opposite case, if agents $\#1, \#2$ were not connected (separated #1), we apply program (4.2) and the related method with the smaller family $A = \{IC_{32}\}$ to find connected packages (w_2, w_3). Then we check (violated or not) IC_{21}: if it is not violated, then the previously found Pareto-optimal w_1 does not change, otherwise, it changes. In the latter case, we again must solve the three-package component $A = \{IC_{32}, IC_{21}\}$. It means that finding w_3 may work in such a way that previously disconnected packages #1 and #2 become connected.

We argue that adding a higher component w_k to the previous locally-optimal partial plan $(w_1, ..., w_{k-1})$ may only increase connectedness, but not break it!

Anyway, the calculations above produce some partial plan — an admissible triple (w_1, w_2, w_3) for three lowest components of the desired solution $(w_1, ..., w_n)$.

5. Further, we proceed in the same way adding agent type #4 (package w_4) to our analysis and checking, whether #4 becomes connected with previous packages, or not. In the latter case, the previous packages remain unchanged, otherwise, they change. When they change, any disconnected (excluded) IC constraints below #4 may become connected into the solution structure.

6. We repeat adding new, higher agent types one-by-one and adjusting the current plan w accordingly until we reach the highest type $\#n$. At each step, adding a higher component w_k to the previous locally-optimal partial plan $(w_1, ..., w_{k-1})$ may only increase connectedness, but not break it. When the connectedness increases, we must recalculate the lower components, otherwise, this is not necessary.

In the end of this algorithm, various outcomes are possible: all types become separated; type 1 and type 2 are linked but other separated; type #1 is separated, types #2 and #3 are linked but other separated; all types can be connected, and so on.

This algorithm gives an exact solution through a finite number of steps, each step solving equations, which are the specific first-order conditions for the related A-structure.

Commenting on the general idea, we observe that starting from the lowest types, we check the connectivity of types. When we meet a new active IC constraint, we check if any previously inactive constraint becomes active. The "impulse" of restrictions goes down through the chain of lower connected types. Indeed, when IC_{i+1} becomes active, the lower-neighbor utility $u_i[w_i]$ decreases, therefore the lower IC_i might force $u_{i-1}[w_{i-1}]$ to decrease also, and so on — this is what we mean by the "impulse".

Why should the algorithm attain the optimal list of constraints A^*? At each step of the process, we maximize the profit (from the partial plan) with the *minimal possible* number of constraints. We always keep the partial plan admissible, always checking if any additional constraints are activated. These considerations are not complete proof, but they support the idea that the solution found by the algorithm should be the global maximum. We suppose that this algorithm reaches the true optimal solution structure A^* and the optimal plan w^*, and that we need not explore any other structures A avoided by this method. We cannot provide more detailed proof of this fact so far.

Conclusion

To summarize, this paper suggests a new model of screening, which is second-degree price discrimination, for situations where consumers are both vertically and horizontally heterogeneous; in their willingness to pay for quality and in their locations in geographical space or space of preferred characteristics of the commodity.

The screening game is reformulated as an optimization program of the seller. The existence of solutions, which are equilibria, is established under typical for the screening literature assumptions. This constrained program being potentially non-convex, the heuristic algorithm is proposed to reduce the search among all possible combinations of active constraints, “solution structures”.

The examples show that solution structures can vary: agent types can be connected by one common chain of IC constraints (called “chain of envy” in the screening literature), or completely separated, or consist of several chains of adjacently-numbered agents.

The important economic feature of such equilibria in spatial screening is that, like in the usual screening, the highest (in each chain) agent type always gets a Pareto-efficient quality, whereas others do not, their quality is distorted downwards. Unlike the usual screening, almost all agents (all except “the farthest consumer”) get some informational rent, their payoffs are higher than their reservation utility.

The most interesting economic extension of this study would be an application of our spatial screening model to oligopoly screening. It promises an explanation of many real-life situations in competition, poorly studied so far.

REFERENCES

1. Chade H., Swinkels J. Screening in vertical oligopolies. *Econometrica,* 2021. Vol. 89, No. 3. P. 1265–1311. DOI: 10.3982/ECTA17016
2. Hotelling H. Stability in competition. *Economic J.*, 1929. Vol. 39, No. 153. P. 41–57. DOI: 10.2307/2224214
3. Katz M. L. Nonuniform pricing with unobservable numbers of purchases. *Rev. Econ. Stud.*, 1984. Vol. 51, No. 3. P. 461–470. DOI: 10.2307/2297434
4. Kokovin S. G., Nahata B. Method of digraphs for multi-dimensional screening. *Ann. Oper. Res.*, 2017. Vol. 253, No. 1. P. 431-451. DOI: 10.1007/s10479-016-2320-3
5. Laffont J-J., Martimort D. *The Theory of Incentives: The Principal–Agent Model.* Princeton: Princeton Univ. Press, 2002. 421 p. DOI: 10.2307/j.ctv7h0rwr
6. Rothschild M., Stiglitz J. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quart. J. Econ.*, 1976. Vol. 90, No. 4. P. 629–649. DOI: 10.2307/1885326
7. Spence A. M. Job market signaling. *Quart. J. Econ.*, 1973. Vol. 87, No. 3. P. 355–374. DOI: 10.2307/1882010
8. Stole L. A. Price discrimination and competition. *Handbook Ind. Organiz.*, 2007. Vol. 3. P. 2221–2299. DOI: 10.1016/S1573-448X(06)03034-2
9. Torbenko A. Linear city models: overview and typology. *J. New Econ. Assoc.*, 2015. Vol. 25, No. 1. P. 12–38. (in Russian) URL: https://ideas.repec.org/a/nea/journl/y2015i25p12-38.html

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 81–86
DOI: 10.15826/umj.2021.1.006

AN ANALOGY OF HAHN-BANACH SEPARATION THEOREM FOR NEARLY TOPOLOGICAL LINEAR SPACES

Madhu Ram

Department of Mathematics, University of Jammu,
Jammu-180006, India
madhuram0502@gmail.com

Abstract: In this paper, we introduce the notion of nearly topological linear spaces and use it to formulate an alternative definition of the Hahn–Banach separation theorem. We also give an example of a topological linear space to which the result is not valid. It is shown that $\mathbb{R}$ with its ordinary topology is not a nearly topological linear space.

Keywords: Hahn–Banach separation theorem, α-open sets, α-compact sets, Nearly topological linear spaces.

1. Introduction

In this paper, all linear spaces are over the field $\mathbb{K} \in \{\mathbb{R},\ \mathbb{C}\}$ unless clear from the context. When we treat $\mathbb{K}$ as a topological space, we mean $\mathbb{K}$ is equipped with its standard topology. For any undefined concepts and terminologies, refer to [9].

Topological linear spaces are intensively studied since they are useful for instance in functional analysis, fixed point theory, equilibrium problems and many others. In functional analysis and fixed point theory, there are many popular theorems which are proven for topological linear spaces like Schauder-Tychonoff fixed point theorem, Hahn Banach separation theorem, etc (for example, see [1, 9, 10]). This paper acquires its inspiration from the following result which is very popular in Functional Analysis and other related branches of Science (see [1, 5, 9, 10], for example) and some papers of its applications (see [2–5, 8], for example):

Theorem 1 (Hahn-Banach Separation Theorem). *Let $\mathfrak{a}$, $\mathfrak{b}$ be disjoint, non-empty convex sets in a topological linear space L.*

(a) *If $\mathfrak{a}$ is open, then there exist a continuous linear functional $\varphi\colon L \to \mathbb{K}$ and $\lambda \in \mathbb{R}$ such that $\operatorname{Re}\varphi(x) < \lambda \le \operatorname{Re}\varphi(y)$, for all $x \in \mathfrak{a}$, $y \in \mathfrak{b}$.*

(b) *If $\mathfrak{a}$ is compact, $\mathfrak{b}$ is closed, and L is locally convex, then there exist a continuous linear functional $\varphi\colon L \to \mathbb{K}$ and $\alpha,\ \beta \in \mathbb{R}$ such that $\operatorname{Re}\varphi(x) < \alpha < \beta < \operatorname{Re}\varphi(y)$, for all $x \in \mathfrak{a}$, $y \in \mathfrak{b}$.*

A natural question to ask is: Is Theorem 1 still valid if L is not a topological linear space? We exhibit that there is a partial answer to this question for a different class of topological linear spaces.

Definition 1. *Let L be a linear space and $\mathfrak{a}$ a subset of L. Then $\mathfrak{a}$ is called*

(1) *convex if $\forall\, x, y \in \mathfrak{a}$, and $\forall\, \alpha, \beta \ge 0$ such that $\alpha + \beta = 1$, $\alpha x + \beta y \in \mathfrak{a}$;*

(2) *absorbing if for each* $x \in L$, $\exists\, r > 0$ *such that* $\forall\, \lambda \in \mathbb{K}$ *with* $|\lambda| \le r$, *we have* $\lambda x \in \mathfrak{a}$;

(3) *balanced if* $\forall\, x \in \mathfrak{a}$, *and* $\forall\, \lambda \in \mathbb{K}$ *with* $|\lambda| \le 1$, *we have* $\lambda x \in \mathfrak{a}$.

Definition 2. *Let L be a linear space and $\mathfrak{c}$ be a non-empty subset of L which is absorbing. The Minkowski (or gauge) functional of $\mathfrak{c}$ is a function, $p\colon L \to \mathbb{R}$, defined as*

$$p(x) = \inf\{\lambda > 0 \colon x \in \lambda\mathfrak{c}\}.$$

Lemma 1 [9, Theorem 1.35]. *Suppose $\mathfrak{c}$ is a convex absorbing set in a vector space L. Then*

(1) $p(x + y) \le p(x) + p(y)$;

(2) $p(\lambda x) = \lambda p(x)$ *if* $\lambda \ge 0$;

(3) p *is a semi-norm if* $\mathfrak{c}$ *is balanced.*

A subset $\mathfrak{a}$ of a topological space X is called α-open [7] if $\mathfrak{a} \subseteq \mathrm{Int}\,(\mathrm{Cl}\,(\mathrm{Int}\,(\mathfrak{a})))$. The complement of an α-open set is called α-closed set. The class of α-open sets of a given topological space X forms a topology on X and it is denoted by $\Im^{\alpha}$. In the following, for given a topological space X we write the corresponding topological space $(X,\ \Im^{\alpha})$ by X^{α}. A subset $\mathfrak{a}$ of X is called α-compact [6] if every cover of $\mathfrak{a}$ by α-open sets of X has a finite subcover.

Note that every open set in a topological space is α-open, every closed set in a topological space is α-closed, and every α-compact set in a topological space is compact, but the converse of these implications is not true in general.

Example 1. Consider the topological space $(X,\ \Im)$ where $X = \mathbb{R}$, and $\Im$ is the usual topology on $\mathbb{R}$. Let

$$\mathfrak{a} = \Big\{x \in \mathbb{R} : -1 < x < 1,\ x \ne \frac{1}{n},\ n \in \mathbb{N}\Big\},$$

where $\mathbb{N}$ denotes the set of positive integers. Then $\mathfrak{a}$ is α-open set in $\mathbb{R}$ which is not open. Further let

$$\mathfrak{b} = \Big\{\frac{1}{n} \colon n \in \mathbb{N}\Big\},$$

then $\mathfrak{b}$ is not closed set in $\mathbb{R}$ but it is α-closed set.

2. The main results

Definition 3. *We call a pair $(L,\ \Im)$ (or simply, L if no confusion arises) nearly topological linear space if:*

- *L is a linear space;*
- *$\Im$ is a topology on L, with*
  (1) *for each α-open set W of L containing the vector sum $x + y$ with $x,\ y \in L$, there exist α-open sets U and V of L containing x and y, respectively such that $U + V \subseteq W$, and*
  (2) *for each α-open set W of L containing the scalar product λx with $x \in L$ and $\lambda \in \mathbb{K}$, there exist an open set U of $\mathbb{K}$ containing λ and an α-open set V of L containing x such that $UV \subseteq W$.*

From this definition, we have immediately:

Remark 1. A nearly topological linear space is not necessarily a topological linear space. Conversely, $\mathbb{R}$ with its usual topology is a topological linear space which is not a nearly topological linear space because for α-open set

$$W = \left\{x \in \mathbb{R} : -1 < x < 1,\ x \neq \frac{1}{n},\ x \neq -\frac{1}{m},\ m, n \in \mathbb{N}\right\}$$

which visibly contains $0 = 0.0$, there do not exist any open set U in $\mathbb{K}$ containing 0 and α-open set V in L containing 0 such that $UV \subseteq W$.

For a nearly topological linear space L, consider the mappings,

$$\sigma_x\colon L^\alpha \to L^\alpha \quad \text{defined as} \quad \sigma_x(y) = x + y,$$
$$\pi_\lambda\colon L^\alpha \to L^\alpha \quad \text{defined as} \quad \pi_\lambda(x) = \lambda x; \quad x, y \in L, \quad \lambda \in \mathbb{K}.$$

Theorem 2. *For a nearly topological linear space L, σ_x and π_λ are continuous.*

P r o o f. Follows from Definition 3. □

For a nearly topological linear space L, we denote by Z_0 the class of α-open sets of L containing the zero vector of L.

Theorem 3. *In a nearly topological linear space L, the following statements are valid*:

(a) *Every $\mathfrak{c} \in Z_0$ is absorbing and balanced.*

(b) *If in addition, $\mathfrak{c} \in Z_0$ is convex, then the Minkowski functional p of $\mathfrak{c}$ is a semi-norm and the set*

$$\{x \in L\colon p(x) < 1\} = \mathfrak{c}.$$

P r o o f. (a) Since $0 = 0.0$, there exist an open set $\mathfrak{u}$ in $\mathbb{K}$ containing 0, and $\mathfrak{v} \in Z_0$ such that $\mathfrak{u}\mathfrak{v} \subseteq \mathfrak{c}$. Then there exist a real $\epsilon > 0$ and an open disk $\mathbb{D}_\epsilon$ with center 0 and radius ϵ such that $\mathbb{D}_\epsilon\mathfrak{v} \subseteq \mathfrak{c}$. By Theorem 2, $\pi_{1/\lambda}$ is continuous, so the set $\mathfrak{a} = \mathbb{D}_\epsilon\mathfrak{v} \in Z_0$. Clearly, $\mathfrak{a}$ is balanced. Next, by Definition 3, we have that for any element $x \in L$, there exists an open set $\mathfrak{u}$ in $\mathbb{K}$ containing 0 s.t. $\mathfrak{u}x \subseteq \mathfrak{c}$. Then there exist a real $r > 0$ and an open disk $\mathbb{D}_r$ with center 0 and radius r such that $\mathbb{D}_r x \subseteq \mathfrak{c}$, showing that $\mathfrak{c}$ is absorbing.
(b) Follows from Lemma 1. □

Theorem 4. *Suppose $\mathfrak{a}$, $\mathfrak{b}$ are disjoint sets in a nearly topological linear space L. If $\mathfrak{a}$ is an α-compact set in L, $\mathfrak{b}$ is an α-closed set in L, then there exists a symmetric set $\mathfrak{u} \in Z_0$ such that $(\mathfrak{a} + \mathfrak{u}) \cap (\mathfrak{b} + \mathfrak{u}) = \emptyset$.*

P r o o f. Let $x \in \mathfrak{a}$ be an element. By Definition 3, there are $\mathfrak{u}_1, \mathfrak{u}_2 \in Z_0$ such that

$$(x + \mathfrak{u}_1 + \mathfrak{u}_2) \cap \mathfrak{b} = \emptyset.$$

Consider,

$$\mathfrak{u} = \mathfrak{u}_1 \cap \mathfrak{u}_2 \cap (-\mathfrak{u}_1) \cap (-\mathfrak{u}_2).$$

Since π_λ is continuous, $\mathfrak{u} \in Z_0$. Consequently, there is a symmetric set $\mathfrak{u}_x \in Z_0$ such that

$$(x + \mathfrak{u}_x + \mathfrak{u}_x + \mathfrak{u}_x) \cap \mathfrak{b} = \emptyset \Rightarrow (x + \mathfrak{u}_x + \mathfrak{u}_x) \cap (\mathfrak{b} + \mathfrak{u}_x) = \emptyset.$$

In a similar vein, we obtain a family

$$\mho = \{x + \mathfrak{u}_x : x \in \mathfrak{a}\}$$

of sets. By Theorem 2, $x + \mathfrak{u}_x$ is α-open set in L. Therefore, for some positive integer n, we have

$$\mathfrak{a} \subseteq \bigcup_{i=1}^{n} (x_i + \mathfrak{u}_{x_i}), \quad x_i \in \mathfrak{a} \quad \text{for all} \quad i = 1, 2, \ldots, n.$$

Let

$$\nu = \bigcap_{i=1}^{n} \mathfrak{u}_{x_i}.$$

Then $\nu \in Z_0$, $\nu = -\nu$, and

$$(\mathfrak{a} + \nu) \cap (\mathfrak{b} + \nu) = \emptyset$$

also. □

Theorem 5. *Suppose $\mathfrak{a}$, $\mathfrak{b}$ are disjoint, non-empty convex sets in a nearly topological linear space L.*

(a) *If $\mathfrak{a}$ is α-open, then there is a linear continuous map $\varphi\colon L^\alpha \to \mathbb{K}$ such that $\operatorname{Re}\varphi(x) < \operatorname{Re}\varphi(y)$, for every $x \in \mathfrak{a}$ and for every $y \in \mathfrak{b}$.*

(b) *If $\mathfrak{a}$ is α-compact, $\mathfrak{b}$ is α-closed and for every $\mathfrak{c} \in Z_0$, there exists a convex set $\mathfrak{c}_0 \in Z_0$ such that $\mathfrak{c}_0 \subseteq \mathfrak{c}$, then there exist a linear continuous map $\varphi\colon L^\alpha \to \mathbb{K}$, $\lambda \in \mathbb{R}$ and an $\epsilon > 0$ such that $\operatorname{Re}\varphi(x) < \lambda < \lambda + \epsilon < \operatorname{Re}\varphi(y)$, for every $x \in \mathfrak{a}$ and for every $y \in \mathfrak{b}$.*

P r o o f. (a) We have two cases.
Case I: $\mathbb{K} = \mathbb{R}$. Fix $x_0 \in \mathfrak{a}$, $y_0 \in \mathfrak{b}$. Let

$$\mathfrak{c} = \mathfrak{a} - \mathfrak{b} + y_0 - x_0.$$

Then $\mathfrak{c}$ is convex set in L, with $\mathfrak{c} \in Z_0$. Let p be the Minkowski functional of $\mathfrak{c}$. By Theorem 3, p is semi-norm on L. Since

$$\mathfrak{a} \cap \mathfrak{b} = \emptyset, \quad y_0 - x_0 = w \notin \mathfrak{c}$$

and so $p(w) \geq 1$.

Consider the linear subspace $M = \mathbb{R}w$ of L and define $\psi\colon M \to \mathbb{R}$ by $\psi(tw) = t$. Evidently, ψ is a linear functional on M s.t.

$$\psi(y) \leq p(y), \quad \forall y \in M.$$

By Hahn–Banach extension theorem, there is a linear functional φ on L s.t.

$$\varphi|_M = \psi \quad \text{and} \quad \varphi(y) \leq p(y), \quad \forall y \in L.$$

Now, for sufficiently small $\epsilon > 0$, take $\mathfrak{u} = (\epsilon\mathfrak{c}) \cap (-\epsilon\mathfrak{c})$. By Theorem 2, $\mathfrak{u} \in Z_0$, and for every $x \in \mathfrak{u}$, $\pm x \in \epsilon\mathfrak{c}$, giving us $\epsilon^{-1}(\pm x) \in \mathfrak{c}$. By Theorem 3, $p(\pm x) < \epsilon$. That is, $|\varphi(x)| < \epsilon$ for all $x \in \mathfrak{u}$.

Next, since for every $x \in \mathfrak{a}$, $y \in \mathfrak{b}$, $\varphi(x - y) < 0$ so we have

$$\varphi(x) \leq \lambda \leq \varphi(y), \quad \text{forall} \quad x \in \mathfrak{a}, \quad y \in \mathfrak{b},$$

where $\lambda = \sup\{\varphi(x) : x \in \mathfrak{a}\}$.

Suppose there exists some $a_0 \in \mathfrak{a}$ s.t. $\varphi(a_0) = \lambda$. By the continuity of the map

$$\mathbb{R} \ni \alpha \;\mapsto\; a_0 + \alpha w \in L^\alpha$$

we have a real number $\epsilon > 0$ such that

$$a_0 + \alpha w \in \mathfrak{a}, \quad \text{for each} \quad \alpha \in \mathbb{R} \quad \text{satisfying} \quad |\alpha| \leq \epsilon.$$

In particular, $a_0 + \epsilon w \in \mathfrak{a}$, showing that $\lambda + \epsilon \leq \lambda$, which is impossible.

Case II: $\mathbb{K} = \mathbb{C}$. The above case gives us a linear continuous function $\varphi\colon L^\alpha \to \mathbb{R}$ with the requisite properties. Then considering the function $\psi(\varsigma) = \varphi(\varsigma) - i\varphi(i\varsigma)$ is the required function, where $i = \sqrt{-1}$.

(b) By Theorem 4, there exists a set $\mathfrak{u} \in Z_0$ such that $(\mathfrak{a} + \mathfrak{u}) \cap \mathfrak{b} = \emptyset$. Then part (a) indicates that there are a continuous linear function $\varphi\colon L^\alpha \to \mathbb{K}$, and $\lambda \in \mathbb{R}$ such that

$$\operatorname{Re}\varphi(x) < \lambda \leq \operatorname{Re}\varphi(y), \quad \text{for every} \quad x \in \mathfrak{a} + \mathfrak{u}, \quad y \in \mathfrak{b}.$$

Since $\mathfrak{a}$ is compact proper subset of $\mathfrak{a} + \mathfrak{u} \subseteq L^\alpha$, $\operatorname{Re}\varphi(\mathfrak{a})$ is compact proper subset of $\operatorname{Re}\varphi(\mathfrak{a} + \mathfrak{u})$. Thus, there exists $\lambda > 0$ such that

$$\operatorname{Re}\varphi(x) < \lambda < \lambda + \epsilon < \operatorname{Re}\varphi(y), \quad \text{for every} \quad x \in \mathfrak{a} \quad \text{and for every} \quad y \in \mathfrak{b}.$$

Whence the proof easily follows. □

Corollary 1. *Suppose $\mathfrak{b}$ is a convex, balanced, α-closed set in a nearly topological linear space L. If $x_0 \in L$, but $x_0 \notin \mathfrak{b}$ and for every $\mathfrak{v} \in Z_0$, there exists a convex set $\mathfrak{u} \in Z_0$ such that $\mathfrak{v} \subseteq \mathfrak{u}$, then there is a continuous linear functional $\varphi\colon L^\alpha \to \mathbb{K}$ such that*

$$|\varphi(x)| \leq 1, \quad \textit{for all} \quad x \in \mathfrak{b}, \quad \textit{and} \quad |\varphi(x_0)| > 1.$$

3. Conclusion

In this paper, we introduced the notion of nearly topological linear spaces and formulated an alternative definition of Hahn–Banach separation theorem by using the notion of α-open sets in topological spaces in the sense of Njastad. It is shown that $\mathbb{R}$ with its ordinary topology is not a nearly topological linear space.

If we endow $\mathbb{C}$, the real linear space of complex numbers with the topology generated by the family of sets of the form

$$D_{r,\epsilon} = \{x + iy\colon x,\ y \in \mathbb{R},\ r - \epsilon < x < r + \epsilon,\ i^2 = -1\}$$

with $r \in \mathbb{R}$ and $\epsilon > 0$, then $\mathbb{C}$ is a nearly topological linear space.

Besides checking the validity of results of topological linear spaces in the field of nearly topological linear spaces, it will be a good contribution finding some more examples of nearly topological linear spaces which satisfy some separation axioms and Theorem 5.

Acknowledgements

The author thanks the referee and the reviewers for their valuable comments/suggestions which improved the quality of the paper.

REFERENCES

1. Amar A. B., Cherif M. A., Mnif M. Fixed-point theory on a Frechet topological vector space. *Int. J. Math. Math. Sci.*, 2011. Vol. 2011. Art. no. 390720. P. 1–9. DOI: 10.1155/2011/390720
2. Deutsch F. R., Maserick P. H. Applications of the Hahn–Banach theorem in approximation theory. *SIAM Rev.*, 1967. Vol. 9, No. 3. P. 516–530. URL: https://www.jstor.org/stable/2027994
3. Deutsch F., Hundal H., Zikatanov L. *Some Applications of the Hahn–Banach Separation Theorem*, 2017. 26 p. arXiv:1712.10250v1 [math.FA]
4. Helton J. W., Klep I., McCullough S. *The Tracial Hahn–Banach Theorem, Polar Duals, Matrix Convex Sets, and Projections of Free Spectrahedra*, 2014. 56 p. arXiv:1407.8198 [math.OA]
5. Luna-Elizarrarás M. E., Perez-Regalado C. O., Shapiro M. On linear functionals and Hahn-Banach theorems for hyperbolic and bicomplex modules. *Adv. Appl. Clifford Algebr.*, 2014. Vol. 24. P. 1105–1129. DOI: 10.1007/s00006-014-0503-z
6. Maheshwari S. N., Thakur S. S. On α-compact spaces. *Bull. Inst. Math. Acad. Sin. (N.S.)*, 1985. Vol. 13, No. 4. P. 341–347.
7. Njåstad O. On some classes of nearly open sets. *Pacific J. Math.*, 1965. Vol. 15, No. 3. P. 961–970.
8. Nörtemann S. The Hahn-Banach theorem for partially ordered totally convex, positively convex and superconvex modules. *Appl. Categ. Structures*, 2002. Vol. 10, p. 417–429. DOI: 10.1023/A:1016390813177
9. Rudin W. *Functional Analysis.* 2nd ed. Singapore: McGraw-Hill Inc., 1991. 448 p.
10. Schaefer H. H. *Topological Vector Spaces.* New York: Springer-Verlag, 1971. 296 p. DOI: 10.1007/978-1-4684-9928-5

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 87–95
DOI: 10.15826/umj.2021.1.007

ON ZYGMUND-TYPE INEQUALITIES CONCERNING POLAR DERIVATIVE OF POLYNOMIALS

Nisar Ahmad Rather

University of Kashmir,
Hazratbal, Srinagar, Jammu and Kashmir 190006, India
dr.narather@gmail.com

Suhail Gulzar

Government College of Engineering and Technology,
Safapora, Ganderbal, Jammu and Kashmir 193504, India
sgmattoo@gmail.com

Aijaz Bhat

University of Kashmir,
Hazratbal, Srinagar, Jammu and Kashmir 190006, India
aijazb824@gmail.com

Abstract: Let $P(z)$ be a polynomial of degree n, then concerning the estimate for maximum of $|P'(z)|$ on the unit circle, it was proved by S. Bernstein that $\|P'\|_\infty \leq n\|P\|_\infty$. Later, Zygmund obtained an L_p-norm extension of this inequality. The polar derivative $D_\alpha[P](z)$ of $P(z)$, with respect to a point $\alpha \in \mathbb{C}$, generalizes the ordinary derivative in the sense that $\lim_{\alpha\to\infty} D_\alpha[P](z)/\alpha = P'(z)$. Recently, for polynomials of the form $P(z) = a_0 + \sum_{j=\mu}^{n} a_j z^j$, $1 \leq \mu \leq n$ and having no zero in $|z| < k$ where $k > 1$, the following Zygmund-type inequality for polar derivative of $P(z)$ was obtained:

$$\|D_\alpha[P]\|_p \leq n\Big(\frac{|\alpha| + k^\mu}{\|k^\mu + z\|_p}\Big)\|P\|_p, \quad \text{where} \quad |\alpha| \geq 1, \quad p > 0.$$

In this paper, we obtained a refinement of this inequality by involving minimum modulus of $|P(z)|$ on $|z| = k$, which also includes improvements of some inequalities, for the derivative of a polynomial with restricted zeros as well.

Keywords: L^p-inequalities, Polar derivative, Polynomials.

1. Zygmund type inequalities for polynomials

Let $\mathcal{P}_n$ denote the space of all complex polynomials of degree at most n. Define

$$\|P\|_p := \left(\frac{1}{2\pi}\int_0^{2\pi} \big|P(e^{i\theta})\big|^p d\theta\right)^{1/p}, \quad 0 < p < \infty.$$

It is well known that the supremum norm satisfies

$$\|P\|_\infty := \max_{|z|=1} |P(z)| = \lim_{p\to\infty} \|P\|_p\,.$$

It is also known [11] that $\lim_{p\to 0}\|P\|_p = \|P\|_0$, where

$$\|P\|_0 := \exp\left(\frac{1}{2\pi}\int_0^{2\pi}\log\left|P(e^{i\theta})\right|d\theta\right).$$

Let $D_\alpha[P](z)$ denote the polar differentiation (see [12]) of a polynomial $P(z)$ of degree n with respect to a complex number α, then

$$D_\alpha[P](z) := nP(z) + (\alpha - z)P'(z).$$

Note that $D_\alpha[P](z)$ is a polynomial of degree at most $n-1$ and it generalizes the ordinary derivative $P'(z)$ of $P(z)$ in the sense that

$$\lim_{\alpha\to\infty}\frac{D_\alpha[P](z)}{\alpha} = P'(z)$$

uniformly with respect to z for $|z| \le R,\ R > 0$.

If $P \in \mathcal{P}_n$, then

$$\|P'\|_p \le n\|P\|_p. \tag{1.1}$$

Inequality (1.1) is due to Zygmund [21] for the case $p \ge 1$. In its proof, he uses M. Riesz's interpolation formula by means of Minkowski's inequality and obtained this inequality as an L_p-norm analogue of Bernstein's inequality (for details see [13] or [20]). A natural question was raised here: whether the restriction on p was indeed necessary? The question remained open for quite a long time despite some partial answers. Finally, it was Arestov [1] came up with some remarkable results which among other things proved that the inequality (1.1) remains valid for $0 < p < 1$ as well. This result is sharp as shown by $P(z) = az^n, a \ne 0$. Arestov [2] also obtained some sharp Bernstein–Zygmund type inequalities for the Szegö composition operators on the set of algebraic polynomials with restrictions on the location of their zeros.

For the class of polynomials $P \in \mathcal{P}_n$ having no zero in $|z| < 1$, inequality (1.1) can be sharpened. In fact, if $P \in \mathcal{P}_n$ and $P(z) \ne 0$ for $|z| < 1$, then

$$\left\|P'\right\|_p \le \frac{n}{\|1+z\|_p}\|P\|_p, \quad p \ge 1. \tag{1.2}$$

Inequality (1.2) is due to De Bruijn [7]. Later Rahman and Schmeisser [16] followed Arestov's technique and proved that this inequality remains true for $0 < p < 1$ as well. The estimates is sharp and equality in (1.2) holds for $P(z) = az^n + b, |a| = |b| \ne 0$.

Govil and Rahman [10] generalized inequality (1.2) and proved that if $P \in \mathcal{P}_n$ does not vanish in $|z| < k$ where $k \ge 1$, then

$$\left\|P'\right\|_p \le \frac{n}{\|k+z\|_p}\|P\|_p, \quad p \ge 1. \tag{1.3}$$

Let $\mathcal{P}_{n,\mu} \subset \mathcal{P}_n$ be a class of lacunary type polynomials

$$P(z) = a_0 + \sum_{j=\mu}^{n} a_j z^j,$$

where $1 \le \mu \le n$.

As a generalization of inequality (1.3), it was shown by Gardner & Weems [8] that if $P \in \mathcal{P}_{n,\mu}$ and $P(z) \ne 0$ for $|z| < k$, $k \ge 1$, then

$$\left\|P'\right\|_p \le \frac{n}{\|k^\mu+z\|_p}\|P\|_p, \quad p > 0. \tag{1.4}$$

Aziz and Rather [5] extended inequality (1.2) to the polar derivative of a polynomial and proved that if $P \in \mathcal{P}_n$ and $P(z)$ does not vanish in $|z| < 1$, then for $\alpha \in \mathbb{C}$ with $|\alpha| \geq 1$, and $p \geq 1$,

$$\|D_\alpha[P]\|_p \leq n\left(\frac{|\alpha|+1}{\|1+z\|_p}\right)\|P\|_p. \tag{1.5}$$

Concerning the concept and properties of the polar derivative refer to [14].

Aziz et. al [6] also obtained an analogue of inequality (1.3) to the polar derivative and proved that if $P \in \mathcal{P}_n$ and $P(z) \neq 0$ for $|z| < k$ where $k \geq 1$, then for $\alpha \in \mathbb{C}$ with $|\alpha| \geq 1$ and $p \geq 1$,

$$\|D_\alpha[P]\|_p \leq n\left(\frac{|\alpha|+k}{\|k+z\|_p}\right)\|P\|_p. \tag{1.6}$$

Rather [17, 18] showed that inequalities (1.5) and (1.6) remain valid for $0 < p < 1$ as well.

Recently, as a generalization of inequality (1.6), Rather et. al [19] proved that if $P \in \mathcal{P}_{n,\mu}$ and $P(z)$ does not vanish in $|z| < k$ where $k \geq 1$, then for $\alpha \in \mathbb{C}$ with $|\alpha| \geq 1$ and $0 \leq p < \infty$,

$$\|D_\alpha[P]\|_p \leq n\left(\frac{|\alpha|+k^\mu}{\|k^\mu+z\|_p}\right)\|P\|_p. \tag{1.7}$$

2. Main results

In this paper, we obtain a refinement of inequality (1.7) by involving the minimum modulus of a polynomial. We prove the following main result.

Theorem 1. *If $P \in \mathcal{P}_{n,\mu}$ and $P(z)$ does not vanish in $|z| < k$ where $k \geq 1$, then for $\alpha \in \mathbb{C}$ with $|\alpha| \geq 1$, $0 \leq p \leq \infty$ and $0 \leq t \leq 1$,*

$$\left\| |D_\alpha[P]| + nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right\|_p \leq n\left(\frac{|\alpha|+k^\mu}{\|z+k^\mu\|_p}\right)\|P\|_p, \tag{2.1}$$

where $m = \min_{|z|=k}|P(z)|$.

Since

$$\frac{nmt(|\alpha|-1)}{1+k^\mu} \geq 0 \quad \text{for} \quad |\alpha| \geq 1,$$

then one can easily observe that

$$\|D_\alpha[P]\|_p \leq \left\| |D_\alpha[P]| + nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right\|_p,$$

and this implies that the Theorem 1 is a refinement of inequality (1.7).

If we divide both sides of inequality (2.1) by $|\alpha|$ and let $|\alpha| \to \infty$, we obtain the following refinement of inequality (1.4).

Corollary 1. *If $P \in \mathcal{P}_{n,\mu}$ and $P(z)$ does not vanish in $|z| < k$ where $k \geq 1$, then for $0 \leq p \leq \infty$,*

$$\left\| |P'| + \frac{nmt}{1+k^\mu}\right\|_p \leq \frac{n}{\|z+k^\mu\|_p}\|P\|_p, \tag{2.2}$$

where $m = \min_{|z|=k}|P(z)|$. The result is best possible as shown by the polynomial

$$P(z) = (z^\mu + k^\mu)^{n/\mu},$$

where μ divides n.

Inequality (2.2) also includes a refinement of (1.3). By taking $k = 1$ and $\mu = 1$ in (2.2), the following improvement of inequality (1.2) follows immediately.

Corollary 2. *If $P \in \mathcal{P}_n$ and $P(z)$ does not vanish in $|z| < 1$ then for $0 \le p \le \infty$,*

$$\left\| |P'| + \frac{nmt}{2} \right\|_p \le \frac{n}{\|1+z\|_p} \|P\|_p, \tag{2.3}$$

where $m = \min_{|z|=1} |P(z)|$. The result is sharp and equality in (2.3) holds for $P(z) = z^n + 1$.

3. Lemmas

For the proof of above theorem, we need the following lemmas.

Lemma 1. *If*

$$P(z) = a_0 + \sum_{j=\mu}^{n} a_j z^j, \quad 1 \le \mu \le n,$$

is a polynomial of degree n having no zeros in $|z| < k$, where $k \ge 1$, then

$$k^\mu |P'(z)| \le |Q'(z)| \quad for \quad |z| = 1,$$

where $Q(z) = z^n \overline{P(1/\overline{z})}$.

The above Lemma 1 is implicit in Qazi [15] and the proof of next lemma is implicit in [9].

Lemma 2. *If $P(z)$ is a polynomial of degree n having no zero in $|z| < k$, $k \ge 1$, then for every $\lambda \in \mathbb{C}$ with $|\lambda| < 1$,*

$$|Q'(z)| \ge |\lambda| mn \quad for \quad |z| = 1,$$

where

$$m = \min_{|z|=k} |P(z)|, \quad Q(z) = z^n \overline{P(1/\overline{z})}.$$

Lemma 3. *If*

$$P(z) = a_0 + \sum_{j=\mu}^{n} a_j z^j, \quad 1 \le \mu \le n,$$

is a polynomial of degree n having no zeros in $|z| < k$, where $k \ge 1$, then for $0 \le t \le 1$,

$$k^\mu |P'(z)| \le |Q'(z)| - mnt \quad for \quad |z| = 1, \tag{3.1}$$

where

$$Q(z) = z^n \overline{P(1/\overline{z})}, \quad m = \min_{|z|=k} |P(z)|.$$

P r o o f. By hypothesis, the polynomial $P(z)$ has no zero in $|z| < k, k \ge 1$. We first show for a given $\lambda \in \mathbb{C}$ with $|\lambda| < 1$, the polynomial $F(z) = P(z) - \lambda m$ does not vanish in $|z| < k$. This is clear if $m = 0$, that is if $P(z)$ has a zero on $|z| = k$. We now suppose that all the zeros of $P(z)$ lie in $|z| > k$, then clearly $m > 0$ so that $m/P(z)$ is analytic in $|z| \le k$ and

$$\left| \frac{m}{P(z)} \right| \le 1 \quad \text{for} \quad |z| = k.$$

Since $m/P(z)$ is not a constant, by the Minimum modulus principle, it follows that

$$m < |P(z)| \quad \text{for} \quad |z| < k. \tag{3.2}$$

Now, if $F(z) = P(z) - \lambda m$ has a zero in $|z| < k$, say at $z = z_0$ with $|z_0| < k$, then

$$P(z_0) - \lambda m = 0.$$

This gives

$$|P(z_0)| = |\lambda m| = |\lambda| m \leq m, \quad \text{where} \quad |z_0| < k,$$

which contradicts (3.2). Hence, we conclude that in any case, the polynomial

$$F(z) = P(z) - \lambda m$$

does not vanish in $|z| < k$, $k \geq 1$, for every $\lambda \in \mathbb{C}$ with $|\lambda| \leq 1$. Applying Lemma 1 to

$$F(z) = P(z) - \lambda m,$$

we get

$$|Q'(z) - \overline{\lambda} mnz^{n-1}| \geq k^{\mu} |P'(z)| \quad \text{for} \quad |z| = 1. \tag{3.3}$$

Now choosing the argument of λ so that on $|z| = 1$,

$$|Q'(z) - \overline{\lambda} mnz^{n-1}| = |Q'(z)| - |\lambda| mn \tag{3.4}$$

which is possible due to lemma 2. By combining (3.3) and (3.4), we obtain

$$|Q'(z)| \geq k^{\mu} |P'(z)| + tmn \quad \text{for} \quad |z| = 1, \tag{3.5}$$

where $t = |\lambda|$ and $0 \leq t < 1$. For the case $t = 1$, the inequality (3.1) follows immediately by letting $t \to 1$ in (3.5) and this completes the proof.

The following lemma is due to Aziz and Rather [3].

Lemma 4. *If A, B and C are non-negative real numbers such that $B + C \leq A$, then for every real number β,*

$$|(A - C) + e^{i\beta}(B + C)| \leq |A + e^{i\beta} B|.$$

Lemma 5 [19]. *If a, b are any two positive real numbers such that $a \geq bc$ where $c \geq 1$, then for any $x \geq 1$, $p > 0$ and $0 \leq \beta < 2\pi$,*

$$(a + bx)^p \int_0^{2\pi} \left| c + e^{i\beta} \right|^p d\beta \leq (c + x)^p \int_0^{2\pi} \left| a + be^{i\beta} \right|^p d\beta.$$

We also need the following lemma due to Aziz and Rather [4].

Lemma 6 [4]. *If $P \in \mathcal{P}_n$ and $Q(z) = z^n \overline{P(1/\overline{z})}$, then for every $p > 0$ and β real , $0 \leq \beta < 2\pi$,*

$$\int_0^{2\pi} \int_0^{2\pi} \left| P'(e^{i\theta}) + e^{i\beta} Q'(e^{i\theta}) \right|^p d\theta d\beta \leq 2\pi n^p \int_0^{2\pi} \left| P(e^{i\theta}) \right|^p d\theta.$$

4. Proof of Theorem 1

P r o o f. By hypothesis $P \in \mathcal{P}_{n,\mu}$ and does not vanish in $|z| < k$, where $k \geq 1$ further if

$$Q(z) = z^n \overline{P(1/\overline{z})},$$

then, by Lemma 3, we have for $|z| = 1$,

$$k^\mu \left|P'(z)\right| \leq \left|Q'(z)\right| - mnt = \left|Q'(z)\right| - mnt \left(\frac{1+k^\mu}{1+k^\mu}\right).$$

Equivalently,

$$k^\mu \left(\left|P'(z)\right| + \frac{mnt}{1+k^\mu}\right) \leq |Q'(z)| - \frac{mnt}{1+k^\mu} \quad \text{for} \quad |z| = 1. \tag{4.1}$$

Setting

$$A = \left|Q'(e^{i\theta})\right|, \quad B = |P'(e^{i\theta})|, \quad C = \frac{mnt}{1+k^\mu}$$

in Lemma 4 we note by (4.1) that

$$B + C \leq k^\mu(B+C) \leq A - C \leq A, \quad \text{since} \quad k \geq 1.$$

Therefore, by Lemma 4 for each real β, we get

$$\left|\left(|Q'(e^{i\theta})| - \frac{mnt}{1+k^\mu}\right) + e^{i\beta}\left(|P'(e^{i\theta})| + \frac{mnt}{1+k^\mu}\right)\right| \leq \left||Q'(e^{i\theta})| + e^{i\beta}|P'(e^{i\theta})|\right|.$$

This implies for each $p > 0$

$$\int_0^{2\pi} \left|F(\theta) + e^{i\beta}G(\theta)\right|^p d\theta \leq \int_0^{2\pi} \left||Q'(e^{i\theta})| + e^{i\beta}|P'(e^{i\theta})|\right|^p d\theta, \tag{4.2}$$

where

$$F(\theta) = |Q'(e^{i\theta})| - \frac{mnt}{1+k^\mu} \quad \text{and} \quad G(\theta) = |P'(e^{i\theta})| + \frac{mnt}{1+k^\mu}. \tag{4.3}$$

Let $P'(\theta) = |P'(\theta)|e^{i\psi}$ and $Q'(\theta) = |Q'(\theta)|e^{i\phi}$, then

$$\int_0^{2\pi} \left|Q'(e^{i\theta})e^{i\beta} + P'(e^{i\theta})\right|^p d\beta = \int_0^{2\pi} \left||Q'(e^{i\theta})|e^{i(\beta+\phi)} + e^{i\psi}|P'(e^{i\theta})|\right|^p d\beta$$
$$= \int_0^{2\pi} \left||Q'(e^{i\theta})|e^{i(\beta+\phi-\psi)} + |P'(e^{i\theta})|\right|^p d\beta.$$

Putting $\beta + \phi - \psi = \Phi$, then we obtain,

$$\int_0^{2\pi} \left|Q'(e^{i\theta})e^{i\beta} + P'(e^{i\theta})\right|^p d\beta = \int_{\phi-\psi}^{2\pi+\phi-\psi} \left||Q'(e^{i\theta})|e^{i\Phi} + |P'(e^{i\theta})|\right|^p d\Phi.$$

Since the function

$$T(\Phi) = |Q'(e^{i\theta})|e^{i\Phi} + |P'(e^{i\theta})|$$

is periodic with period 2π, hence we have

$$\begin{aligned}\int_0^{2\pi}\left|Q'(e^{i\theta})e^{i\beta}+P'(e^{i\theta})\right|^p d\beta &= \int_0^{2\pi}\left||Q'(e^{i\theta})|e^{i\Phi}+|P'(e^{i\theta})|\right|^p d\Phi\\ &= \int_0^{2\pi}\left||Q'(e^{i\theta})|e^{i\beta}+|P'(e^{i\theta})|\right|^p d\beta.\end{aligned} \tag{4.4}$$

Integrating (4.2) both sides with respect to β from 0 to 2π and using (4.4), we get

$$\begin{aligned}\int_0^{2\pi}\int_0^{2\pi}\left|F(\theta)+e^{i\beta}G(\theta)\right|^p d\theta d\beta &\leq \int_0^{2\pi}\int_0^{2\pi}\left||Q'(e^{i\theta})|+e^{i\beta}|P'(e^{i\theta})|\right|^p d\beta d\theta\\ &= \int_0^{2\pi}\int_0^{2\pi}\left|Q'(e^{i\theta})+e^{i\beta}P'(e^{i\theta})\right|^p d\beta d\theta\\ &= \int_0^{2\pi}\int_0^{2\pi}\left|P'(e^{i\theta})+e^{i\beta}Q'(e^{i\theta})\right|^p d\theta d\beta.\end{aligned}$$

By using Lemma 6 this implies,

$$\int_0^{2\pi}\int_0^{2\pi}\left|F(\theta)+e^{i\beta}G(\theta)\right|^p d\theta d\beta \leq 2\pi n^p\int_0^{2\pi}\left|P(e^{i\theta})\right|^p d\theta. \tag{4.5}$$

Now for $|z|=1$, $0\leq t\leq 1$ and $\alpha\in\mathbb{C}$ with $|\alpha|\geq 1$ and using the fact that

$$|nP(z)-zP'(z)|=|Q'(z)|$$

for z with unit modulus, we have

$$\begin{aligned}|D_\alpha[P](e^{i\theta})|+nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right) &= \left|nP(z)+(\alpha-z)P'(z)\right|+nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\\ &\leq |\alpha||P'(z)|+|nP(z)-zP'(z)|+nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\\ &= |\alpha||P'(e^{i\theta})|+|Q'(e^{i\theta})|+nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\\ &= |\alpha|\left(|P'(e^{i\theta})|+\frac{mnt}{1+k^\mu}\right)+\left(|Q'(e^{i\theta})|-\frac{mnt}{1+k^\mu}\right).\end{aligned}$$

By integrating both sides with respect to θ from 0 to 2π, for each $p>0$, we get

$$\begin{aligned}&\int_0^{2\pi}\left||D_\alpha[P](e^{i\theta})|+nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right|^p d\theta\\ &\leq \int_0^{2\pi}\left||\alpha|\left(|P'(e^{i\theta})|+\frac{mnt}{1+k^\mu}\right)+\left(|Q'(e^{i\theta})|-\frac{mnt}{1+k^\mu}\right)\right|^p d\theta.\end{aligned}$$

Multiply both sides by

$$\int_0^{2\pi}|k^\mu+e^{i\beta}|^p d\beta,$$

we obtain

$$\begin{aligned}&\int_0^{2\pi}|k^\mu+e^{i\beta}|^p d\beta\int_0^{2\pi}\left||D_\alpha[P](e^{i\theta})|+nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right|^p d\theta\\ &\leq \int_0^{2\pi}\left||\alpha|\left(|P'(e^{i\theta})|+\frac{mnt}{1+k^\mu}\right)+\left(|Q'(e^{i\theta})|-\frac{mnt}{1+k^\mu}\right)\right|^p d\theta\int_0^{2\pi}|k^\mu+e^{i\beta}|^p d\beta.\end{aligned} \tag{4.6}$$

Further, since $k^\mu \geq 1$, $1 \leq \mu \leq n$, and if

$$a = \left|Q'(e^{i\theta})\right| - \frac{mnt}{1+k^\mu}, \quad b = \left|P'(e^{i\theta})\right| + \frac{mnt}{1+k^\mu}, \quad c = k^\mu, \quad x = |\alpha|\,,$$

then from (4.1) one can observe that $a \geq bc$. Using Lemma 5, we get for every $\alpha \in \mathbb{C}$ with $|\alpha| \geq 1$,

$$\begin{gathered}\left\{\left(|Q'(e^{i\theta})| - \frac{mnt}{1+k^\mu}\right) + |\alpha|\left(|P'(e^{i\theta})| + \frac{mnt}{1+k^\mu}\right)\right\}^p \int_0^{2\pi} |k^\mu + e^{i\beta}|^p d\beta \\ \leq (|\alpha| + k^\mu)^p \int_0^{2\pi} \left|\left(|Q'(e^{i\theta})| - \frac{mnt}{1+k^\mu}\right) + e^{i\beta}\left(|P'(e^{i\theta})| + \frac{mnt}{1+k^\mu}\right)\right|^p d\beta.\end{gathered}$$

Again, integrating both sides with respect to θ from 0 to 2π, we obtain

$$\begin{gathered}\int_0^{2\pi} \left|\left(|Q'(e^{i\theta})| - \frac{mnt}{1+k^\mu}\right) + |\alpha|\left(|P'(e^{i\theta})| + \frac{mnt}{1+k^\mu}\right)\right|^p d\theta \int_0^{2\pi} |k^\mu + e^{i\beta}|^p d\beta \\ \leq (|\alpha| + k^\mu)^p \int_0^{2\pi}\int_0^{2\pi} \left|F(\theta) + e^{i\beta}G(\theta)\right|^p d\beta d\theta,\end{gathered}$$

where $F(\theta)$ and $G(\theta)$ are given by (4.3). Using this in inequality (4.6), we get

$$\begin{gathered}\int_0^{2\pi} |k^\mu + e^{i\beta}|d\beta \int_0^{2\pi} \left||D_\alpha[P](e^{i\theta})| + nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right|^p d\theta \\ \leq (|\alpha| + k^\mu)^p \int_0^{2\pi}\int_0^{2\pi} |F(\theta) + e^{i\beta}G(\theta)|^p d\beta d\theta.\end{gathered} \tag{4.7}$$

By using (4.5) in (4.7), we obtain for each $p > 0$ and $|\alpha| \geq 1$

$$\int_0^{2\pi} |k^\mu + e^{i\beta}|d\beta \int_0^{2\pi} \left||D_\alpha[P](e^{i\theta})| + nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right|^p d\theta \leq (|\alpha| + k^\mu)^p 2\pi n^p \int_0^{2\pi} \left|P(e^{i\theta})\right|^p d\theta.$$

Equivalently,

$$\begin{gathered}\left(\frac{1}{2\pi}\int_0^{2\pi} \left||D_\alpha[P](e^{i\theta})| + nmt\left(\frac{|\alpha|-1}{1+k^\mu}\right)\right|^p d\theta\right)^{1/p} \\ \leq \frac{n(|\alpha| + k^\mu)}{\left(1/(2\pi)\int_0^{2\pi} |k^\mu + e^{i\beta}|d\beta\right)^{1/p}} \left(\frac{1}{2\pi}\int_0^{2\pi} \left|P(e^{i\theta})\right|^p d\theta\right)^{1/p},\end{gathered}$$

which immediately leads to (2.1) for $0 < p < \infty$ and the cases $p = 0$ and $p = \infty$ follow by respectively taking the limits $p \to 0^+$ and $p \to \infty$. This completes the proof of Theorem 1. $\square$

Acknowledgement

We are thankful to the referee for useful comments and suggestions.

REFERENCES

1. Arestov V. V. On integral inequalities for trigonometric polynomials and their derivatives. *Math. USSR-Izv.*, 1982. Vol. 18, No. 1. P. 1–17. DOI: 10.1070/IM1982v018n01ABEH001375
2. Arestov V. V. Integral inequalities for algebraic polynomials with a restriction on their zeros. *Anal. Math.*, 1991. Vol. 17, P. 11–20. DOI: 10.1007/bf02055084

3. Aziz A., Rather N. A. L^p inequalities for polynomials. *Glas. Math.*, 1997. Vol. 32, No. 1. P. 39–43.
4. Aziz A., Rather N. A. Some Zygmund type L^q inequalities for polynomials. *J. Math. Anal. Appl.*, 2004. Vol. 289, No. 1. P. 14–29. DOI: 10.1016/S0022-247X(03)00530-4
5. Aziz A., Rather N. A. On an inequality concerning the polar derivative of a polynomial. *Proc. Math. Sci.*, 2007. Vol. 117. P. 349–357. DOI: 10.1007/s12044-007-0030-0
6. Aziz A, Rather N. A., Aliya Q. L_q norm inequalities for the polar derivative of a polynomial. *Math. Inequal. Appl.*, 2008. Vol. 11. P. 283–296. DOI: 10.7153/mia-11-20
7. De Bruijn N G. Inequalities concerning polynomials in the complex domain. *Indag. Math. (N.S.)*, 1947. Vol. 9, No. 5. P. 1265–1272.
8. Gardner R., Weems A. A Bernstein type L^p inequality for a certain class of polynomials. *J. Math. Anal. Appl.*, 1998. Vol. 219. P. 472–478.
9. Govil N. K. On the growth of polynomials. *J. Inequal. Appl.*, 2002. Vol. 7, No. 5. P. 623–631.
10. Govil N. K., Rahman Q. I. Functions of exponential type not vanishing in a half-plane and related polynomials. *Trans. Amer. Math. Soc.*, 1969. Vol. 137, P. 501–517. DOI: 10.1090/S0002-9947-1969-0236385-6
11. Mahler K. An application of Jensen's formula to polynomials. *Mathematika*, 1960. Vol. 7, No. 2. P. 98–100. DOI: 10.1112/S0025579300001637
12. Marden M. *Geometry of Polynomials.* Math. Surveys, Amer. Math. Soc., 1989. 243 p.
13. Milovanovic G. V., Mitrinovic D. S., Rassias Th. *Topics in Polynomials: Extremal properties, Inequalituies, Zeros.* Singapore: World Scientific, 1994. 836 p. DOI: 10.1142/1284
14. Pólya G.,Szegö G. *Aufgaben und lehrsätze aus der Analysis.* Springer-Verlag, Berlin, 1925. 353 p. (in German)
15. Qazi M. A. On the maximum modulus of polynomials. *Proc. Amer. Math. Soc.*, 1992. Vol. 115. P. 237–243. DOI: 10.1090/S0002-9939-1992-1113648-1
16. Rahman Q. I., Schmeisser G. L^p inequalities for polynomials. *J. Approx. Theory*, 1998. Vol. 53, No. 1. P. 26–32. DOI: 10.1016/0021-9045(88)90073-1
17. Rather N. A., Some integral inequalities for the polar derivative of a polynomial. *Math. Balkanica (N.S.)*, 2008. Vol. 22, No. 3–4. P. 207–216.
18. Rather N. A. L^p inequalities for the polar derivative of a polynomial. *J. Inequal. Pure Appl. Math.*, 2008. Vol. 9, No. 4. Art. no. 103, P. 1–10.
19. Rather N. A., Iqbal A., Hyun G. H. Integral inequalities for the polar derivative of a polynomial. *Nonlinear Funct. Anal. Appl.*, 2018. Vol. 23, No. 2. P. 381–393.
20. Schaeffer A. C. Inequalities of A. Markoff and S.Bernstein for polynomials and related functions. *Bull. Amer. Math. Soc.*, 1941. Vol. 47. P. 565–579. DOI: 10.1090/S0002-9904-1941-07510-5
21. Zygmund A. A remark on conjugate series. *Proc. Lond. Math. Soc. (3)*, 1932. Vol. s2-34, No. 1. P. 392–400. DOI: 10.1112/plms/s2-34.1.392

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 96–101

DOI: 10.15826/umj.2021.1.008

DEFINITE INTEGRAL OF LOGARITHMIC FUNCTIONS AND POWERS IN TERMS OF THE LERCH FUNCTION[1]

Robert Reynolds†, **Allan Stauffer**††

Department of Mathematics and Statistics,
York University, 4700 Keele Street, Toronto, Canada

†milver@my.yorku.ca, ††stauffer@yorku.ca

Abstract: A family of generalized definite logarithmic integrals given by

$$\int_0^1 \frac{\left(x^{im}(\log(a)+i\log(x))^k + x^{-im}(\log(a)-i\log(x))^k\right)}{(x+1)^2}dx$$

built over the Lerch function has its analytic properties and special values listed in explicit detail. We use the general method as given in [5] to derive this integral. We then give a number of examples that can be derived from the general integral in terms of well known functions.

Keywords: Entries of Gradshteyn and Ryzhik, Lerch function, Knuth's Series.

1. Introduction

In connection with logarithmic integrals, the authors have the opportunity to evaluate integrals of the form

$$\int_0^1 \frac{\left(x^{im}(\log(a)+i\log(x))^k + x^{-im}(\log(a)-i\log(x))^k\right)}{(x+1)^2}dx \tag{1.1}$$

in terms of the Lerch function. We chose this integral as it forms the general case for some integrals published in the Tables of Gradshteyn and Rhyzik. It yields some very interesting special cases in terms of Euler–Mascheroni constant (γ), and a pair of Zeta function values $\zeta(1/2)$ and $\zeta(-1/2)$. The constant $\zeta(1/2)$ is used to calculate Knuth's Series and a new integral representation for this constant is derived. The Lerch function is also used in the Bose–Einstein condensation for an exponential density of states function [4]. We also provide formal derivations for some definite integrals in [3] not previously listed in current literature along with new definite integrals in terms of special functions. In our case the constants in the equation (1.1) are general complex numbers subject to the restrictions given below. The derivations follow the method used by us in [5]. The generalized Cauchy's integral formula is given by

$$\frac{y^k}{k!} = \frac{1}{2\pi i}\int_C \frac{e^{wy}}{w^{k+1}}dw. \tag{1.2}$$

This method involves using a form of equation (1.2) then multiply both sides by a function, then takes a definite integral of both sides. This yields a definite integral in terms of a contour integral. Then we multiply both sides of equation (1.2) by another function and take the infinite sum of both sides such that the contour integral of both equations are the same.

[1]This research is supported by NSERC Canada under Grant 504070. The authors confirm there are no conflicts of interest.

2. Definite integral of the contour integral

We use the method given in [5]. The contour integral is over $\alpha = m + w$. Here the contour is in the upper left quadrant with $\Im(\alpha) < 0$ and going round the origin with zero radius. Using a generalization of Cauchy's integral formula we first replace y by $\log(a) + ix$ then multiply by e^{mxi} for the first equation and then y by $\log(a) - ix$ and multiply by e^{-mxi} to get the second equation followed by replacing x by $\log(x)$. Then we add these two equations, followed by multiplying both sides by $1/2(x+1)^2$ to get the equality

$$\frac{\left(x^{im}(\log(a) + i\log(x))^k + x^{-im}(\log(a) - i\log(x))^k\right)}{2(x+1)^2 k!} = \frac{1}{2\pi i}\int_C \frac{a^w w^{-k-1}\cos(\alpha\log(x))}{(x+1)^2}d\alpha. \tag{2.1}$$

Next we take the definite integral of equation (2.1) over $x \in [0,1]$ to get the following relations

$$\begin{aligned}
&\int_0^1 \frac{\left(x^{im}(\log(a) + i\log(x))^k + x^{-im}(\log(a) - i\log(x))^k\right)}{2(x+1)^2 k!}dx \\
&= \frac{1}{2\pi i}\int_0^1\int_C \frac{a^w w^{-k-1}\cos(\alpha\log(x))}{(x+1)^2}d\alpha dx \\
&= \frac{1}{2\pi i}\int_C\int_0^1 \frac{a^w w^{-k-1}\cos(\alpha\log(x))}{(x+1)^2}dx d\alpha \\
&= \frac{1}{2\pi i}\int_C \frac{1}{2}\pi(m+w)a^w w^{-k-1}\operatorname{csch}(\pi(m+w))dw
\end{aligned} \tag{2.2}$$

from equation (3.883.1) in [3] where the logarithmic function is defined in equation (4.1.2) in [1]. The integral is valid for a, k and m complex and $\Im(\alpha) < 0$.

3. Infinite sum of the contour integral

In this section we will again use the generalized Cauchy's integral formula to derive equivalent contour integrals. First we replace y by $\log(a)+\pi(2y+1))$ and multiply both sides by $-m\pi e^{\pi m(2y+1)}$ to get

$$-\frac{\pi^{k+1} m e^{\pi m(2y+1)}\left(\log(a)/\pi + 2y + 1\right)^k}{k!} = -\frac{1}{2\pi i}\int_C \pi\alpha w^{-k-1}\exp(w\log(a) + \pi\alpha(2y+1))d\alpha.$$

Next we take the infinite sum over $y \in [0,\infty)$ simplify the left-hand in terms of the Lerch function side to get

$$\begin{aligned}
&-\frac{2^k\pi^{k+1}e^{\pi m}m}{k!}\Phi\left(e^{2m\pi}, -k, \frac{\log(a)+\pi}{2\pi}\right) \\
&= -\frac{1}{2\pi i}\sum_{y=0}^{\infty}\int_C \pi m w^{-k-1}(\exp(w\log(a) + \pi\alpha(2y+1)))d\alpha \\
&= -\frac{1}{2\pi i}\int_C\sum_{y=0}^{\infty} \pi m w^{-k-1}(\exp(w\log(a) + \pi\alpha(2y+1)))d\alpha \\
&= \frac{1}{2\pi i}\int_C \frac{1}{2}\pi m a^w w^{-k-1}\operatorname{csch}(\pi\alpha)d\alpha.
\end{aligned} \tag{3.1}$$

Next we derive the second contour integral by replacing k by $k-1$ and dropping the linear factor m in equation (3.1) to get

$$-\frac{2^{k-1}\pi^k e^{\pi m}\Phi\left(e^{2m\pi},1-k,(\log(a)+\pi)/2\pi\right)}{(k-1)!}=\frac{1}{2\pi i}\int_C \frac{1}{2}\pi a^w w^{-k}\operatorname{csch}(\pi\alpha)d\alpha$$

from (1.232.3) in [3], where csch(x) = $i$csc(ix) from (4.5.10) in [1] and $\Im(\alpha)<0$ for the sum to converge.

We use (9.550) and (9.556) in [3] where $\Phi(z,s,v)$ is the Lerch function which is a generalization of the Hurwitz Zeta and polylogarithm functions.

The Lerch function has a series representation given by

$$\Phi(z,s,v)=\sum_{n=0}^{\infty}(v+n)^{-s}z^n,$$

where $|z|<1$, $v\neq 0,-1,..$ and is continued analytically by its integral representation given by

$$\Phi(z,s,v)=\frac{1}{\Gamma(s)}\int_0^{\infty}\frac{t^{s-1}e^{-vt}}{1-ze^{-t}}dt=\frac{1}{\Gamma(s)}\int_0^{\infty}\frac{t^{s-1}e^{-(v-1)t}}{e^t-z}dt,$$

where $\mathrm{Re}\,(v)>0$, or $|z|\leq 1$, $z\neq 1$, $\mathrm{Re}\,(s)>0$, or $z=1$, $\mathrm{Re}\,(s)>1$.

4. Definite integral in terms of the Lerch function

Since the right-hand sides of equation (2.2) and (3.1) are equivalent we can equate the left-hand sides simplifying the factorials to get

$$\begin{aligned}&\int_0^1\frac{\left(x^{im}(\log(a)+i\log(x))^k+x^{-im}(\log(a)-i\log(x))^k\right)}{(x+1)^2}dx\\&=(2\pi)^k\left(-e^{\pi m}\right)\left(k\Phi\left(e^{2m\pi},1-k,\frac{\log(a)+\pi}{2\pi}\right)+2\pi m\Phi\left(e^{2m\pi},-k,\frac{\log(a)+\pi}{2\pi}\right)\right).\end{aligned}\tag{4.1}$$

5. Derivation of entry (4.325.3) in [3]

In this section will derive an integral representation for the Riemann zeta function. Using equation (4.1) setting $m=0$, $a=1$ and simplifying the left-hand side we get

$$\int_0^1\frac{\log^k(1/x)}{(x+1)^2}dx=2^{-k}(2^k-2)\zeta(k)\Gamma(k+1).\tag{5.1}$$

This formula is equivalent to applying integration by parts to equation (1.12.5) in [2].

Next we take the partial derivative with respect to k of equation (5.1) simplifying to get

$$\begin{aligned}&\int_0^1\frac{\log(\log(1/x))\log^k(1/x)}{(x+1)^2}dx\\&=2^{-k}\Gamma(k+1)\Big((2^k-2)\zeta'(k)+\zeta(k)\big((2^k-2)\psi^{(0)}(k+1)+\log(4)\big)\Big).\end{aligned}\tag{5.2}$$

Next we set $k=0$ and simplify to get

$$\int_0^1\log(\log(1/x))\frac{dx}{(x+1)^2}=\frac{1}{2}(\log(\pi/2)-\gamma)$$

from [7, p. 236].

6. Derivation of special case in terms of $\zeta(-1/2)$

Using equation (5.2) and setting $k = -1/2$ we get

$$\int_0^1 \frac{\log(\log(1/x))}{\sqrt{\log(1/x)}} \frac{dx}{(x+1)^2}$$
$$= \sqrt{\pi}\Big(2\sqrt{2}\zeta(-1/2)\log(2) - (2\sqrt{2}-1)(\zeta'(-1/2) + \zeta(-1/2)\psi^{(0)}(1/2))\Big).$$

7. Derivation of special cases in terms of $\zeta(1/2)$

Using equation (5.2) and setting $k = 1/2$ we get

$$\int_0^1 \sqrt{\log(1/x)}\log\left(\log(1/x)\right)\frac{dx}{(x+1)^2}$$
$$= \frac{\sqrt{\pi}}{8}\zeta(1/2)\Big(8 - 8\sqrt{2} + 2(\sqrt{2}-1)\gamma + \pi + \sqrt{2}(\log(64/\pi^2) - \pi) + \log(\pi^2/4)\Big)$$

from [7, p. 236].

8. Derivation of a special case involving combinations of rational functions of $\log(x)$ and powers

8.1. Definite integral in terms of the hypergeometric and Lerch functions

Setting $k = -1$ and replacing a by e^a in (4.1) we derive one equation by replacing m by ip and a second equation by replacing m by $-ip$ then subtracting the two equations and simplifying to get

$$\int_0^1 \frac{(x^p - x^{-p})\log(x)}{(a^2 + \log^2(x))}\frac{dx}{(x+1)^2}$$
$$= \frac{1}{4\pi(a+\pi)}\Bigg(ie^{-i\pi p}(a+\pi)\Phi\left(e^{-2ip\pi}, 2, \frac{a+\pi}{2\pi}\right) - i(a+\pi)e^{i\pi p}\Phi\left(e^{2ip\pi}, 2, \frac{a+\pi}{2\pi}\right)$$
$$- 4\pi^2 pe^{-i\pi p}\left({}_2F_1\left(1, \frac{a+\pi}{2\pi}; \frac{1}{2}\left(\frac{a}{\pi}+3\right); e^{-2ip\pi}\right) + e^{i\pi p}\,{}_2F_1\left(1, \frac{a+\pi}{2\pi}; \frac{1}{2}\left(\frac{a}{\pi}+3\right); e^{2ip\pi}\right)\right)\Bigg)$$

from equation (9.559) in [3] and where $\Re(a) > 0$. This is a new entry for Table 4.282 in [3].

8.2. Definite integral in terms of the Lerch functions

Setting $k = -2$ and replacing a by e^a in (4.1) we derive one equation by replacing m by ip and a second equation by replacing m by $-ip$ then subtracting the two equations and simplifying to get

$$\int_0^1 \frac{(x^p - x^{-p})\log(x)}{\left(a^2 + \log^2(x)\right)^2}\frac{dx}{(x+1)^2} = \frac{1}{8\pi^2 a}\Bigg(- \pi pe^{-i\pi p}\Phi\left(e^{-2ip\pi}, 2, \frac{a+\pi}{2\pi}\right)$$
$$+ ie^{-i\pi p}\Phi\left(e^{-2ip\pi}, 3, \frac{a+\pi}{2\pi}\right) - e^{i\pi p}\left(\pi p\Phi\left(e^{2ip\pi}, 2, \frac{a+\pi}{2\pi}\right) + i\Phi\left(e^{2ip\pi}, 3, \frac{a+\pi}{2\pi}\right)\right)\Bigg),$$

where $\Re(a) > 0$. This is a new entry for Table 4.282 in [3].

9. Derivation of a special case of combinations involving powers of the logarithm and other powers

9.1. Derivation in terms of the hyperbolic tangent and Lerch functions

Setting $k = -1$ and $a = 1$ in (4.1) we derive one equation by replacing m by ip and a second equation by replacing m by $-ip$ then subtracting the two equations and simplifying to get

$$\int_0^1 \frac{(x^p - x^{-p})}{(x+1)^2} \frac{dx}{\log(x)} = -p \tanh^{-1}(\cos(\pi p)) + \frac{i}{4\pi}\left(e^{-i\pi p}\Phi\left(e^{-2ip\pi}, 2, \frac{1}{2}\right) - e^{i\pi p}\Phi\left(e^{2ip\pi}, 2, \frac{1}{2}\right)\right)$$

from equations (9.121.27) and (9.559) in [3]. This is a new entry for Table 4.283 in [3].

9.2. Derivation in terms of Lerch function

Setting $a = 1$ in (4.1) we derive one equation by replacing m by ip and a second equation by replacing m by $-ip$ then subtracting the two equations and simplifying the logarithmic functions on the left-hand side to get

$$\int_0^1 \log^k(1/x)\left(x^p - x^{-p}\right)(x+1)^2 dx = i2^{k-1}\pi^k \csc\left(\frac{\pi k}{2}\right)\Bigg(-ke^{-i\pi p}\Phi\left(e^{-2ip\pi}, 1-k, \frac{1}{2}\right) + 2i\pi p e^{-i\pi p}\Phi\left(e^{-2ip\pi}, -k, \frac{1}{2}\right) + e^{i\pi p}\left(k\Phi\left(e^{2ip\pi}, 1-k, \frac{1}{2}\right) + 2i\pi p\Phi\left(e^{2ip\pi}, -k, \frac{1}{2}\right)\right)\Bigg).$$

This is a new entry for Table 4.272 in [3].

10. Discussion

In this paper we have derived a new integral representation for $\zeta(1/2)$ the value of which is apparently unknown in terms of known constants. We were able to derive an efficient method for evaluating Knuth's series using this integral representation. We also derived a new integral representation for evaluating $\zeta(-1/2)$. We have dealt with a similar integral in the paper "A Definite Integral Involving the Logarithmic Function in Terms of the Lerch Function" [6]. The present paper should be seen as an extension of these results.

11. Conclusion

In this paper, we have presented a novel method for deriving some interesting definite integrals using contour integration. The results presented were numerically verified for both real and imaginary and complex values of the parameters in the integrals using Mathematica by Wolfram.

REFERENCES

1. Abramowitz M., Stegun I. A. (Eds.) *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, 9th printing. New York, Dover, 1972. 1046 p.
2. Erdélyi A., Magnus W., Oberhettinger F., Tricomi F. G. *Higher Transcendental Functions*. Vol. 1. New York–Toronto–London: McGraw-Hill Book Company Inc., 1953. 316 p.
3. Gradshteyn I. S., Ryzhik I. M. *Table of Integrals, Series and Products*, 7 ed. Academic Press, 2007. 1171 p.
4. Momeni D. Bose–Einstein condensation for an exponential density of states function and Lerch zeta function. *Phys. A*, 2020. Vol. 541, Art. No. 123264. 9 p. DOI: 10.1016/j.physa.2019.123264

5. Reynolds R., Stauffer A. A method for evaluating definite integrals in terms of special functions with examples. *Int. Math. Forum*, 2020. Vol. 15, No. 5. P. 235–244. DOI: 10.12988/imf.2020.91272
6. Reynolds R., Stauffer A. A Definite Integral Involving the Logarithmic Function in Terms of the Lerch Function. *Mathematics*, 2019. Vol. 7, No. 12. Art. No. 1148. 5 p. DOI: 10.3390/math7121148
7. Whittaker E. T., Watson G. N. *A Course of Modern Analysis*, 4th ed. Cambridge, England: Cambridge University Press, 1996. 608 p.

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 102–108

DOI: 10.15826/umj.2021.1.009

THE VERTEX DISTANCE COMPLEMENT SPECTRUM OF SUBDIVISION VERTEX JOIN AND SUBDIVISION EDGE JOIN OF TWO REGULAR GRAPHS

Ann Susa Thomas

Department of Mathematics, St Thomas College,
Kozhencherry-689641, Kerala, India
anns11thomas@gmail.com

Sunny Joseph Kalayathankal

Jyothi Engineering College,
Cheruthuruthy, Thrissur-679531, Kerala, India
sjkalayathankal@jecc.ac.in

Joseph Varghese Kureethara

Department of Mathematics, Christ University,
Bangalore-560029, Karnataka, India
frjoseph@christuniversity.in

Abstract: The vertex distance complement (VDC) matrix C, of a connected graph G with vertex set consisting of n vertices, is a real symmetric matrix $[c_{ij}]$ that takes the value $n - d_{ij}$ where d_{ij} is the distance between the vertices v_i and v_j of G for $i \neq j$ and 0 otherwise. The vertex distance complement spectrum of the subdivision vertex join, $G_1 \dot{\bigvee} G_2$ and the subdivision edge join $G_1 \underline{\bigvee} G_2$ of regular graphs G_1 and G_2 in terms of the adjacency spectrum are determined in this paper.

Keywords: Distance matrix, Vertex distance complement spectrum, Subdivision vertex join, Subdivision edge join.

1. Introduction

Spectral graph theory deals with the study of the eigenvalues of various matrices associated with graphs. Initially, the spectrum of the adjacency matrix of a graph was studied. Collatz and Sinogowitz initiated the exploration of this topic in 1957 [2]. Since then spectral theory of graphs is an active research area [1, 3].

In this paper, we consider the matrix derived from a type of distance matrix, viz., vertex distance complement (VDC) matrix. The VDC spectra of some classes of graphs are found in [8, 9]. The VDC matrix C of a graph G [7] is defined as follows

$$C = \begin{cases} n - d_{ij}, & i \neq j, \\ 0, & i = j, \end{cases}$$

where d_{ij} is the distance between the vertices v_i and v_j of G and n denotes the number of vertices of G.

The subdivision graph $S(G)$ of a graph G is obtained by inserting a new vertex of degree two in every edge of G. Let $V(G)$ and $I(G)$ denote respectively the existing vertex set and the set of the newly introduced vertices of the subdivision graph $S(G)$ of a graph G. The adjacency spectrum of two joins, $G_1 \dot{\bigvee} G_2$ and $G_1 \underline{\bigvee} G_2$, based on subdivision graph was determined in [4]. The distance spectrum of the same was calculated in [6].

Throughout this article we consider connected simple graphs of diameter at most two. We determine the VDC spectrum of $G_1 \dot{\bigvee} G_2$ and $G_1 \underline{\bigvee} G_2$ when G_1 and G_2 are regular graphs. The eigenvalues of $VDC(G)$ are called the VDC-eigenvalues of G and they form the VDC spectrum of G, denoted by $spec_{VDC}(G)$. We denote J and I as the all-one matrix and identity matrix, respectively, of appropriate orders.

The definitions of the subdivision graphs are as follows.

Definition 1 [4]. *The subdivision-vertex join $G_1 \dot{\bigvee} G_2$ of two vertex disjoint graphs G_1 and G_2 is the graph obtained from $S(G_1)$ and G_2 by joining each vertex of $V(G_1)$ with every vertex of $V(G_2)$.*

Definition 2 [4]. *The subdivision-edge join $G_1 \underline{\bigvee} G_2$ of two vertex disjoint graphs G_1 and G_2 is the graph obtained from $S(G_1)$ and G_2 by joining each vertex of $I(G_1)$ with every vertex of $V(G_2)$.*

The following results are very useful for computing the VDC spectrum.

Lemma 1 [3]. *Let G be an r-regular graph with adjacency matrix A and incidence matrix R. Let $A(L(G))$ denote the adjacency matrix of the line graph $L(G)$ of G. Then,*

$$RR^T = A + rI, \quad R^T R = A(L(G)) + 2I.$$

Also,

$$JR = 2J = R^T J, \quad JR^T = rJ = RJ.$$

Lemma 2 [3]. *Let G be r-regular $(n;m)$ graph with $\operatorname{spec}(G) = \{\lambda_1, \lambda_2, \cdots, \lambda_n\}$. Then*

$$\operatorname{spec}(L(G)) = \begin{cases} 2r-2, & \\ \lambda_i + r - 2, & i = 2, 3, \ldots, n, \\ -2, & m - n \text{ times.} \end{cases}$$

Also, Z is an eigenvector corresponding to the eigenvalue -2 if and only if $RZ = 0$ where R is the incidence matrix of G.

Theorem 1 (Perron–Frobenius). *If all entries of an $n \times n$ matrix are positive, then it has a unique maximal eigenvalue. Its eigenvector has positive entries.*

2. The VDC spectrum of $G_1 \dot{\bigvee} G_2$

Theorem 2. *Let G_i be an r_i regular graph with n_i vertices and m_i edges, for $i = 1, 2$. If $\{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{in_i}\}$ denotes the adjacency spectrum corresponding to the adjacency matrix A_i of G_i, the $\operatorname{spec}_{VDC}(G_1 \dot{\bigvee} G_2)$ consists of*

(i) $2\lambda_{1i} + 2r_1 - n + 2, \quad$ *for* $\quad i = 2, 3, \ldots, n_1;$

(ii) $-n$, *repeated* $m_1 - 1$ *times*;

(iii) $\lambda_{2i} - n + 2$, *for* $i = 2, 3, \ldots, n_2$;

(iv) *the* 3 *roots of the equation*

$$\begin{gathered}
x^3 - (n_1n - 2n_1 + n_2n - 2n_2 + m_1n^2 - 4m_1 + 4r_1 + r_2 - 3n + 4)x^2 \\
-(2n_1n_2n - 3n_1n_2 + n_1m_1 - 2n_1r_1n + 2n_1r_1 - n_1r_2n + 2n_1r_2 + 2n_1n^2 \\
-6n_1n + 4n_1 + 2n_2m_1n - 4n_2m_1 - 4n_2r_1n + 8n_2r_1 + 2n_2n^2 - 6n_2n + 4n_2 - m_1r_2n \\
+4m_1r_2 + 2m_1n^2 - 8m_1n + 4m_1 - 4r_1r_2 + 8r_1n - 8r_1 + 2r_2n - 2r_2 - 3n^2 + 8n - 4)x \\
-(2n_1n_2m_1 - 4n_1n_2r_1n + 4n_1n_2r_1 + 2n_1n_2n^2 - 3n_1n_2n - n_1m_1r_2 \\
+n_1m_1n - 2n_1m_1 + 2n_1r_1r_2n - 2n_1r_1r_2 - 2n_1r_1n^2 + 6n_1r_1n - 4n_1r_1 - n_1r_2n^2 \\
+2n_1r_2n + n_1n^3 - 4n_1n^2 + 4n_1n + 2n_2m_1n^2 - 8n_2m_1 - 4n_2r_1n^2 + 8n_2r_1n + n_2n^3 \\
-4n_2n^2 + 4n_2n - m_1r_2n^2 + 2m_1r_2n + 4m_1r_2 + m_1n^3 - 4m_1n^2 + 8m_1 - 4r_1r_2n \\
+4r_1n^2 - 8r_1n + r_2n^2 - 2r_2n - n^3 + 4n^2 - 4n) = 0,
\end{gathered}$$

where $n = n_1 + m_1 + n_2$.

P r o o f. Given that G_1 and G_2 are regular graphs with regularity r_1 and r_2 respectively. Let R be the incidence matrix of G_1 and $A(L(G_1)$ be the adjacency matrix of the line graph of G_1. The distance matrix of a graph with diameter at most two and adjacency matrix A can be rewritten as $A + 2\bar{A}$ or $2(J - I) - A$ [5].

The subdivision-vertex join $G_1 \dot{\bigvee} G_2$ has $n = n_1 + m_1 + n_2$ vertices. With the proper labeling of vertices, the VDC matrix of $G_1 \dot{\bigvee} G_2$ is a square matrix of order n given by

$$C = \begin{pmatrix} (n-2)(J-I) & (n-3)J + 2R & (n-1)J \\ (n-3)J + 2R^T & (n-4)(J-I) + 2A(L(G_1)) & (n-2)J \\ (n-1)J & (n-2)J & (n-2)(J-I) + A_2 \end{pmatrix}.$$

Let X be an eigenvector corresponding to the eigenvalue $\lambda_{1i} \neq r_1$ of A_1. Using Lemma 1, we note that

$$A(L(G_1))R^TX = (\lambda_{1i} + r_1 - 2)R^TX.$$

Hence, $\lambda_{1i} + r_1 - 2$ are the eigenvalues of $A(L(G_1))$ with an eigenvector R^TX.

By Perron–Frobenius theorem, X and R^TX are orthogonal to the all-one vector J.

Let

$$\Upsilon = \begin{pmatrix} X \\ R^TX \\ 0 \end{pmatrix}.$$

Then,

$$2\lambda_{1i} + 2r_1 - n + 2, \quad i = 2, 3, \ldots, n_1$$

is an eigenvalue of the VDC matrix of $G_1 \dot{\bigvee} G_2$ corresponding to the eigenvector Υ. This is because

$$\begin{pmatrix} (n-2)(J-I) & (n-3)J+2R & (n-1)J \\ (n-3)J+2R^T & (n-4)(J-I)+2A(L(G)) & (n-2)J \\ (n-1)J & (n-2)J & (n-2)(J-I)+A_2 \end{pmatrix} \begin{pmatrix} X \\ R^TX \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} -(n-2)X+2(A_1+r_1I))X \\ 2R^TX-(n-4)R^TX+2A(L(G_1))R^TX \\ 0 \end{pmatrix} = \begin{pmatrix} (2\lambda_{1i}+2r_1-n+2)X \\ (2\lambda_{1i}+2r_1-n+2)R^TX \\ 0 \end{pmatrix}$$
$$= (2\lambda_{1i}+2r_1-n+2) \begin{pmatrix} X \\ R^TX \\ 0 \end{pmatrix}.$$

By a similar reasoning, if Y is an eigenvector of $A(L(G_1))$ corresponding to the eigenvalue $\lambda_{1i}+r_1-2$, for $i=2,3,\dots,n_1$,

$$\Phi = \begin{pmatrix} RY \\ -Y \\ 0 \end{pmatrix}$$

is an eigenvector of VDC matrix of $G_1 \dot{\bigvee} G_2$ corresponding to the eigenvalue $-n$. (Note that the line graph of a regular graph is also regular).

Hence, $-n$ is an eigenvalue of $G_1 \dot{\bigvee} G_2$ repeated n_1-1 times.

Now, -2 is an eigenvalue of $A(L(G_1))$ with multiplicity m_1-n_1. Let Z be an eigenvector of $A(L(G_1))$ corresponding to the eigenvalue -2. Then, by Lemma 2, $RZ=0$ and by Perron–Frobenius theorem, $JZ=0$.

Let

$$\Omega = \begin{pmatrix} 0 \\ Z \\ 0 \end{pmatrix}.$$

Then $-n$ is an eigenvalue of the VDC matrix of $G_1 \dot{\bigvee} G_2$ repeated m_1-n_1 times with an eigenvector Ω. This is because

$$\begin{pmatrix} (n-2)(J-I) & (n-3)J+2R & (n-1)J \\ (n-3)J+2R^T & (n-4)(J-I)+2A(L(G)) & (n-2)J \\ (n-1)J & (n-2)J & (n-2)(J-I)+A_2 \end{pmatrix} \begin{pmatrix} 0 \\ Z \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} 0 \\ -(n-4)Z+2A(L(G_1))Z \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -nZ \\ 0 \end{pmatrix}.$$

In total, $-n$ is an eigenvalue of $G_1 \dot{\bigvee} G_2$ repeated m_1-1 times.

Now, let $\lambda_{2i} \neq r_2$ be an eigenvalue of G_2 with an eigenvector W. Since G_2 is regular, $JW=0$. Hence

$$\Psi = \begin{pmatrix} 0 \\ 0 \\ W \end{pmatrix}$$

is an eigenvector of the VDC matrix of $G_1 \dot{\bigvee} G_2$ corresponding to the eigenvalue $\lambda_{2i}-n+2$, for $i=2,3,\dots n_2$. Thus, we have obtained $n_1+m_1+n_2-3$ eigenvalues.

The remaining three eigenvalues are to be determined. We note that all the eigenvectors constructed so far, are orthogonal to

$$\begin{pmatrix} J \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ J \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 0 \\ J \end{pmatrix}.$$

The remaining three eigenvectors are spanned by these three vectors and is of the form

$$\Theta = \begin{pmatrix} \alpha J \\ \beta J \\ \gamma J \end{pmatrix}.$$

for some $(\alpha, \beta, \gamma) \neq (0, 0, 0)$.
Thus, if ρ is an eigenvalue of the VDC matrix with an eigenvector Θ, then from $C\Theta = \rho\Theta$, we can see that the remaining three eigenvalues are obtained from the matrix

$$\begin{pmatrix} (n-2)(n_1-1) & (n-3)m_1 + 2r_1 & (n-1)n_2 \\ (n-3)n_1 + 4 & n(m_1-1) - 4(m_1 - r_1) & (n-2)n_2 \\ (n-1)n_1 & (n-2)m_1 & (n-2)(n_2-1) + r_2 \end{pmatrix}.$$

Thus we determine the VDC spectrum of $G_1 \dot{\bigvee} G_2$. □

3. The VDC spectrum of $G_1 \underline{\bigvee} G_2$

In this section we present the VDC spectrum of $G_1 \underline{\bigvee} G_2$.

Theorem 3. *Let G_i be r_i regular graph with n_i vertices and m_i edges, for $i = 1, 2$. If $\{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{in_i}\}$ denotes the adjacency spectrum corresponding to the adjacency matrix A_i of G_i, then, the* $\text{spec}_{VDC}(G_1 \underline{\bigvee} G_2)$ *consists of*

(i) $\lambda_{1i} + 3 \pm \sqrt{(\lambda_{1i}+1)^2 + 4(\lambda_{1i} + r_1)} - n,$ *for* $i = 2, 3, \ldots, n_1$;

(ii) $-n + 2,$ *repeated* $m_1 - n_1$ *times*;

(iii) $\lambda_{2i} - n + 2,$ *for* $i = 2, 3, \ldots, n_2$;

(iv) *the* 3 *roots of the equation*

$$\begin{gathered} x^3 - (n_1n - 4n_1 + n_2n - 2n_2 + m_1n - 2m_1 + 2r_1 + r_2 - 3n + 8)x^2 \\ -(2n_1n_2n - 4n_1n_2 + n_1m_1 + 2n_1r_1n - 6n_1r_1 - n_1r_2n + 4n_1r_2 + 2n_1n^2 - 12n_1n + 16n_1 \\ +2n_2m_1n - 3n_2m_1 - 2n_2r_1n + 4n_2r_1 + 2n_2n^2 - 10n_2n + 12n_2 - 2m_1r_1n + 4m_1r_1 + 2m_1n^2 \\ -6m_1n - m_1r_2n - 2r_1r_2 + 4r_1n + 2m_1r_2 + 2r_2n - 6r_2 - 3n^2 + 16n - 20)x \\ -(2n_1n_2m_1 + 4n_1n_2r_1n - 8n_1n_2r_1 + 8n_1n_2 + 2n_1r_1n^2 - 8n_1r_2 - 16n_1 \\ -4n_2m_1r_1n + 6n_2m_1r_1 - 3n_2m_1n - 16n_2 + 2m_1r_1r_2n - 4m_1r_1r_2 - 2m_1r_1n^2 \\ +8m_1r_1n - 8m_1r_1 - 4r_1r_2 + 2n_1n_2n^2 - 8n_1n_2n - n_1m_1r_2 + n_1m_1n - 2n_1m_1 \\ -2n_1r_1r_2n + 6n_1r_1r_2 + 2n_1r_1n^2 - 10n_1r_1n + 12n_1r_1 - n_1r_2n^2 + 6n_1r_2n \\ +n_1n^3 - 8n_1n^2 + 20n_1n + 2n_2m_1n^2 - 4n_2m_1 - 2n_2r_1n^2 + 8n_2r_1 + n_2n^3 \\ -8n_2n^2 + 20n_2n - m_1r_2n^2 + 2m_1r_2n + 4m_1r_2 + m_1n^3 - 4m_1n^2 + 8m_1 - 2r_1r_2n \\ +2r_1n^2 - 8r_1 + r_2n^2 - 6r_2n - n^3 + 8n^2 - 20n + 8r_2 + 16) = 0. \end{gathered}$$

where $n = n_1 + m_1 + n_2$.

P r o o f. Given that G_1 and G_2 are regular graphs with regularity r_1 and r_2 respectively. Let R be the incidence matrix of G_1. $G_1 \underline{\bigvee} G_2$ has $n = n_1 + m_1 + n_2$ vertices. With the proper labeling of vertices, the VDC matrix of $G_1 \underline{\bigvee} G_2$ of order n is given by

$$C = \begin{pmatrix} (n-4)(J-I)+2A_1 & (n-3)J+2R & (n-2)J \\ (n-3)J+2R^T & (n-2)(J-I) & (n-1)J \\ (n-2)J & (n-1)J & (n-2)(J-I)+A_2 \end{pmatrix}.$$

Let $\lambda_{1i} \neq r_1$ be an eigenvalue of A_1 with an eigenvector X. By Perron–Frobenius theorem, X is orthogonal to the all-one vector J.

Let us test the condition under which

$$\Upsilon = \begin{pmatrix} tX \\ R^T X \\ 0 \end{pmatrix}$$

is an eigenvector of the given VDC matrix.

If Υ is an eigenvector of the VDC matrix of $G_1 \underline{\bigvee} G_2$ corresponding to the eigenvalue η, then $C\Upsilon = \eta\Upsilon$ implies

$$\begin{pmatrix} (n-4)(J-I)+2A_1 & (n-3)J+2R & (n-2)J \\ (n-3)J+2R^T & (n-2)(J-I) & (n-1)J \\ (n-2)J & (n-1)J & (n-2)(J-I)+A_2 \end{pmatrix} \begin{pmatrix} tX \\ R^T X \\ 0 \end{pmatrix} = \eta \begin{pmatrix} tX \\ R^T X \\ 0 \end{pmatrix}$$

i. e.,

$$-(n-4)t + 2t\lambda_{1i} + 2\lambda_{1i} + 2r_1 = \eta t \tag{3.1}$$

and

$$2t - (n-2) = \eta. \tag{3.2}$$

Substituting the value of η from equation (3.2) in equation (3.1), we get a quadratic equation in t as

$$t^2 - (1+\lambda_{1i})t - (\lambda_{1i} + r_1) = 0$$

Hence

$$t = \frac{(1+\lambda_{1i}) \pm \sqrt{(1+\lambda_{1i})^2 + 4(\lambda_{1i}+r_1)}}{2}.$$

Thus corresponding to each eigenvalue $\lambda_{1i} \neq r_1$ of A_1, we get two VDC eigenvalues $\eta = 2t + 2 - n$ of $G_1 \underline{\bigvee} G_2$ and hence a total of $2(n_1 - 1)$ VDC eigenvalues are obtained.

Now, -2 is an eigenvalue of $A(L(G_1))$ with multiplicity $m_1 - n_1$. Let Z be an eigenvector of $A(L(G_1))$ with eigenvalue -2. Then, by Lemma 2, $RZ = 0$.

However,

$$\Omega = \begin{pmatrix} 0 \\ Z \\ 0 \end{pmatrix}$$

is an eigenvector of the VDC matrix of $G_1 \underline{\bigvee} G_2$ corresponding to the eigenvalue $-n+2$.

Let $\lambda_{2i} \neq r_2$ be an eigenvalue of G_2 with an eigenvector W. Then,

$$\Psi = \begin{pmatrix} 0 \\ 0 \\ W \end{pmatrix}$$

is an eigenvector of the VDC matrix of $G_1 \underline{\bigvee} G_2$ corresponding to the eigenvalue $\lambda_{2i} - n + 2$, for $i = 2, 3, \ldots n_2$.

Thus, we have obtained $n_1 + m_1 + n_2 - 3$ eigenvalues.

Next, we will determine the remaining three eigenvalues. We note that all the eigenvectors constructed are orthogonal to

$$\begin{pmatrix} J \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ J \\ 0 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 0 \\ 0 \\ J \end{pmatrix}.$$

The remaining three eigenvectors are spanned by these three vectors and is of the form

$$\Theta = \begin{pmatrix} \alpha J \\ \beta J \\ \gamma J \end{pmatrix}$$

for some $(\alpha, \beta, \gamma) \neq (0, 0, 0)$. Thus, if ρ is an eigenvalue of C with an eigenvector Θ then from $C\Theta = \rho\Theta$, we can see that the remaining three eigenvalues are obtained from the matrix

$$\begin{pmatrix} (n-4)(n_1-1)+2r_1 & (n-3)m_1+2r_1 & (n-2)n_2 \\ (n-3)n_1+4 & (n-2)(m_1-1) & (n-1)n_2 \\ (n-2)n_1 & (n-1)m_1 & (n-2)(n_2-1)+r_2 \end{pmatrix}.$$

$\square$

4. Conclusion

In this paper we have computed the Vertex Distance Complement Spectrum of Subdivision Vertex Join, $G_1 \dot{\bigvee} G_2$, and Subdivision Edge Join, $G_1 \underline{\bigvee} G_2$ of regular graphs G_1 and G_2. The work can be extended to graphs with diameter greater than two, graphs that are not regular etc. It is worth exploring the nature of the spectrum of graphs with arbitrary subdivisions.

REFERENCES

1. Brouwer A. E., Haemers W. H. *Spectra of Graphs*. New York: Springer, 2011. 250 p. DOI: 10.1007/978-1-4614-1939-6
2. Collatz L. V., Sinogowitz U. Spektren endlicher grafen. In: *Abh. Math. Semin. Univ. Hambg.*, 1957. Vol. 21, No. 1. P. 63–77. DOI: 10.1007/BF02941924 (in German)
3. Cvetković D. M., Doob M., Sachs H. *Spectra of Graphs — Theory and Application*. New York: Academic Press, 1980. 368 p.
4. Indulal G. Spectrum of two new joins of graphs and infinite families of integral graphs. *Kragujevac J. Math.*, 2012. Vol. 36, No. 1. P. 133–139.
5. Indulal G., Gutman I., Vijayakumar A. On distance energy of graphs. *MATCH Commun. Math. Comput. Chem.,* 2008. Vol. 60, No. 2. P. 461–472.
6. Indulal G., Scaria D. C., Liu X. The distance spectrum of the subdivision vertex join and subdivision edge join of two regular graphs. *Discrete Math. Lett.,* 2019. Vol. 1. P. 36–41.
7. Janežič D., Miličević A., Nikolić S., Trinajstić N. *Graph-Theoretical Matrices in Chemistry*. Florida: CRC Press, 2015. 174 p.
8. Varghese R. P., Susha D. Vertex distance complement spectra of regular graphs and its line graphs. *Int. J. Appl. Math. Anal. Appl.,* 2017. Vol. 12, No. 2. P. 221–231.
9. Varghese R. P., Susha D. Vertex distance complement spectra of some graphs. *Ann. Pure Appl. Math.,* 2018. Vol. 16, No. 1. P. 69–80. DOI: 10.22457/apam.v16n1a9

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 109–119

DOI: 10.15826/umj.2021.1.010

MODIFIED PROXIMAL POINT ALGORITHM FOR MINIMIZATION AND FIXED POINT PROBLEM IN CAT(0) SPACES

Godwin Chidi Ugwunnadi

Department of Mathematics, University of Eswatini,
Private Bag 4, Kwaluseni, Eswatini

Department of Mathematics and Applied Mathematics,
Sefako Makgato Health Science University,
P.O. Box 94, Pretoria 0204, South Africa

ugwunnadi4u@yahoo.com

Abstract: In this paper, we study modified-type proximal point algorithm for approximating a common solution of a lower semi-continuous mapping and fixed point of total asymptotically nonexpansive mapping in complete CAT(0) spaces. Under suitable conditions, some strong convergence theorems of the proposed algorithms to such a common solution are proved.

Keywords: Proximal point algorithm, Total asymptotically nonexpansive mapping, Fixed point, $\triangle$ convergence, Strong convergence, CAT(0) space.

1. Introduction

In recent years, much attention has been given to develop several iterative methods including the *proximal point algorithms* (PPA) which was suggested by Martinet [26] for solving convex optimization problems which was extensively developed by Rockafellar [28] in the context of monotone variational inequalities. The main idea of this method consists of replacing the initial problem with a sequence of regularized problems, so that each particular auxiliary problem can be solved by one of the well-known algorithms. Quiet number of different method of proximal point algorithm have been proposed and studied from the classical linear spaces such as Euclidean spaces, Hilbert spaces, and Banach spaces to the setting of manifolds (see [5, 6, 13, 18, 20, 26, 28]).
Recently, the classical proximal point algorithms have been extended from linear spaces such as Hilbert spaces or Banach spaces to the setting of nonlinear version.

In 2013, Bačák [6] introduced the PPA in a CAT(0) space (X, d) as follows: $x_1 \in X$ and

$$x_{n+1} = \arg\min_{y\in X}\Big(f(y) + \frac{1}{2\lambda_n} d^2(y, x_n)\Big), \quad \forall\ n \geq 1,$$

where $\lambda_n > 0,\ \forall\ n \geq 1$. It was shown that if f has a minimizer and

$$\sum_{n=1}^{\infty} \lambda_n = \infty,$$

then the sequence $\{x_n\}$ $\triangle-$converges to its minimizer [5].

It is a known fact that iterative methods for finding fixed points of nonexpansive mappings have received vast investigations due to its extensive applications in a variety of applied areas of inverse problem, partial differential equations, image recovery, and signal processing; see [2, 5, 8, 15, 21] and the references therein.

Fixed-point theory in CAT(0) spaces was first studied by Kirk [22, 23]. He showed that every nonexpansive (single-valued) mapping defined on a bounded closed convex subset of a complete CAT(0) space always has a fixed point. Since then, the fixed-point theory for single-valued and multivalued mappings in CAT(0) spaces has been rapidly developed.

Recently, Suparatulatorn et al. [29] presented a new modified proximal point algorithm for solving the minimization of a convex function and the fixed points of nonexpansive mappings in CAT(0) spaces. Chang et al. [12] proved some strong convergence theorems of the PPA to a common fixed point of asymptotically nonexpansive mappings and to minimizers of a convex function in CAT(0) spaces.

Let C be a nonempty subset of a complete CAT(0) space X and T a mapping from C into itself. Then, a point $x \in C$ is called a fixed point of T if $Tx = x$. We denote by $F(T)$ the set of all the fixed points of T. A mapping T from C into itself is said to be:

(N) nonexpansive if $d(Tx, Ty) \leq d(x, y)$ for all $x, y \in C$;

(AN) asymptotically nonexpansive, if there is a sequence $\{u_n\} \subseteq [0, \infty)$ with $u_n \to 0$ as $n \to \infty$ such that
$$d(T^n x, T^n y) \leq (1 + u_n) d(x, y), \quad \forall n \geq 1, \quad x, y \in C;$$

(UL) uniformly L-Lipschitzian, if there exists a constant $L > 0$ such that
$$d(T^n x, T^n y) \leq L d(x, y), \quad \forall n \geq 1, \quad x, y \in C.$$

The concept of total asymptotically nonexpansive mappings was first introduced by Alber et al. [1]. A mapping $T : C \to C$ is said to be total asymptotically nonexpansive mapping if there exists nonnegative sequences $\{\mu_n\}, \{\nu_n\}$ with $\mu_n \to 0$, $\nu_n \to 0$ as $n \to \infty$ and a strictly increasing continuous function $\zeta : [0, \infty) \to [0, \infty)$ with $\zeta(0) = 0$ such that
$$d(T^n x, T^n y) \leq d(x, y) + \nu_n \zeta(d(x, y)) + \mu_n, \quad \forall n \geq 1, \quad x, y \in C.$$

Remark 1. From the definitions, it is known that each nonexpansive mapping is asymptotically nonexpansive mapping with sequence $\{u_n = 0\}$, and each asymptotically nonexpansive mapping is $(\{\mu_n\}, \{\nu_n\}, \zeta)$-total asymptotically nonexpansive mapping with $\mu_n = 0$, $\nu_n = u_n$, $\forall n \geq 1$ and $\zeta(t) = t$, $t \geq 0$. But the opposite may not be true for each of them in a general sense. Furthermore, every asymptotically nonexpansive mapping is a uniformly L-Lipschitzian mapping with
$$L = \sup_{n \geq 1} (1 + u_n).$$

Motivated and inspired by the above works, in this paper, we study a modified algorithm for proximal point and fixed point of total asymptotically nonexpansive mapping in CAT(0) space. Strong convergence of this algorithm is proved. Our method of proof is different from the method in Chang et al. [12].

2. Preliminaries

Let (X, d) be a metric space and $x, y \in X$ with $d(x, y) = l$. A geodesic path from x to y is an isometry $c : [0, l] \to X$ such that $c(0) = x$ and $c(l) = y$. The image of a geodesic path is called a *geodesic segment*. A metric space X is a (uniquely) *geodesic space*, if every two points of X are joined by only one geodesic segment. A geodesic triangle $\triangle(x_1, x_2, x_3)$ in a geodesic space X consists of three points x_1, x_2, x_3 of X and three geodesic segments joining each pair of vertices. A comparison triangle of a geodesic triangle $\triangle(x_1, x_2, x_3)$ is the triangle $\bar{\triangle}(x_1, x_2, x_3) := \triangle(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ in the Euclidean space $\mathbb{R}^2$ such that

$$d(x_i, x_j) = d_{\mathbb{R}^2}(\bar{x}_i, \bar{y}_j), \quad \forall i, j = 1, 2, 3.$$

A geodesic space X is a CAT(0) space, if for each geodesic triangle $\triangle(x_1, x_2, x_3)$ in X and its comparison triangle $\bar{\triangle} := \triangle(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ in $\mathbb{R}^2$, the CAT(0) inequality $d(x, y) \leq d_{\mathbb{R}^2}(\bar{x}, \bar{y})$ is satisfied for all $x, y \in \triangle$ and $\bar{x}, \bar{y} \in \bar{\triangle}$.

A thorough discussion of these spaces and their important role in various branches of mathematics are given in [9, 10]. Let $x, y \in X$ and $\lambda \in [0, 1]$, we write $\lambda x \oplus (1-\lambda)y$ for the unique point z in the geodesic segment joining from x to y such that

$$d(z, x) = (1-\lambda)d(x, y) \quad \text{and} \quad d(z, y) = \lambda d(x, y).$$

We also denote by $[x, y]$ the geodesic segment joining from x to y, that is,

$$[x, y] = \big\{\lambda x \oplus (1-\lambda)y : \lambda \in [0, 1]\big\}.$$

A subset C of a CAT(0) space is convex if $[x, y] \subseteq C$ for all $x, y \in C$.

Berg and Nikolaev [7] introduced the concept of an inner product-like notion (quasilinearization) in complete CAT(0) spaces to resolve these difficulties as follows:

Let denote a pair $(a, b) \in X \times X$ by $\overrightarrow{ab}$ and call it a vector. The quasilinearization is a map $\langle ., . \rangle : (X \times X) \times (X \times X) \to \mathbb{R}$ defined by

$$\langle \overrightarrow{ab}, \overrightarrow{cd} \rangle = \frac{1}{2}\Big(d^2(a, d) + d^2(b, c) - d^2(a, c) - d^2(b, d)\Big), \quad \forall a, b, c, d \in X. \tag{2.1}$$

It is easily seen that $\langle \overrightarrow{ab}, \overrightarrow{cd} \rangle = \langle \overrightarrow{cd}, \overrightarrow{ab} \rangle, \langle \overrightarrow{ab}, \overrightarrow{cd} \rangle = -\langle \overrightarrow{ba}, \overrightarrow{cd} \rangle$ and $\langle \overrightarrow{ax}, \overrightarrow{cd} \rangle + \langle \overrightarrow{xb}, \overrightarrow{cd} \rangle = \langle \overrightarrow{ab}, \overrightarrow{cd} \rangle$ for all $a, b, c, d \in X$. We say that X satisfies the Cauchy-Schwarz inequality if

$$\langle \overrightarrow{ab}, \overrightarrow{cd} \rangle \leq d(a, b)d(c, d)$$

for all $a, b, c, d \in X$. It is known that a geodesically connected metric space is a CAT(0) space if and only if it satisfies the Cauchy–Schwarz inequality (see [7]).

Lemma 1 [16]. *Let X be a* CAT(0) *space, $x, y, z \in X$ and $\lambda \in [0, 1]$. Then*

$$d(\lambda x \oplus (1-\lambda)y, z) \leq \lambda d(x, z) + (1-\lambda)d(y, z).$$

Lemma 2 [16]. *Let X be a* CAT(0) *space, $x, y, z \in X$ and $\lambda \in [0, 1]$. Then*

$$d^2(\lambda x \oplus (1-\lambda)y, z) \leq \lambda d^2(x, z) + (1-\lambda)d^2(y, z) - \lambda(1-\lambda)d^2(x, y).$$

Lemma 3 [14]. *Let X be a* CAT(0) *space, $x, y, z \in X$ and $\lambda \in [0, 1]$. Then*

$$d^2(\lambda x \oplus (1-\lambda)y, z) \leq \lambda^2 d^2(x, z) + (1-\lambda)^2 d^2(y, z) + 2\lambda(1-\lambda)\langle \overrightarrow{xz}, \overrightarrow{yz} \rangle.$$

Let $\{x_n\}$ be a bounded sequence in a complete CAT(0) space X. For $x \in X$, we set

$$r(x, \{x_n\}) = \limsup_{n\to\infty} d(x, x_n).$$

The asymptotic radius $r(\{x_n\})$ of $\{x_n\}$ is given by

$$r(\{x_n\}) = \inf\big\{r(x, \{x_n\}) : x \in X\big\},$$

and the asymptotic center $A(\{x_n\})$ of $\{x_n\}$ is the set

$$A(\{x_n\}) = \big\{x \in X : r(x, \{x_n\}) = r(\{x_n\})\big\}.$$

It is well known that in a CAT(0) space $A(\{x_n\})$ consists of exactly one point (see [15, Proposition 7]).

Lemma 4 [24]. *Every bounded sequence in a complete* CAT(0) *space always has a* $\triangle-$*convergent subsequence.*

Lemma 5 [19]. *Let X be a complete* CAT(0) *space, $\{x_n\}$ be a sequence in X and $x \in X$. Then $\{x_n\}$ $\triangle-$converges to x if and only if* $\limsup_{n\to\infty}\langle\overrightarrow{xx_n}, \overrightarrow{xy}\rangle \le 0$ *for all $y \in C$.*

A function $f : C \to (-\infty, \infty]$ defined on a convex subset C of a CAT(0) space is convex if, for any geodesic

$$[x, y] := \{\gamma_{x,y}(\lambda) : 0 \le \lambda \le 1\} := \{\lambda x \oplus (1-\lambda)y : 0 \le \lambda \le 1\}$$

joining $x, y \in C$, the function $f \circ \gamma$ is convex, i.e.

$$f(\gamma_{x,y}(\lambda)) := f(\lambda x \oplus (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y).$$

For examples of convex functions in CAT(0), see [12]. For any $\lambda > 0$, define the Moreau–Yosida resolvent of f in CAT(0) space X as

$$J_\lambda(x) = \arg\min_{y\in X}\Big[f(y) + \frac{1}{2\lambda}d^2(y, x)\Big], \quad \forall\, x \in X.$$

Let $f : X \to (-\infty, \infty]$ be a proper convex and lower semi-continuous function. It is shown in [3] that the set $F(J_\lambda)$ of fixed points of the resolvent associated with f coincides with the set $\operatorname{argmin}_{y\in X} f(y)$ of minimizers of f. Also for any $\lambda > 0$, the resolvent J_λ of f is nonoexpansive [17].

Lemma 6 (Sub-differential inequality [4]). *Let (X, d) be a complete* CAT(0) *space and $f : X \to (-\infty, \infty]$ be proper convex and lower semi-continuous. Then, for all $x, y \in X$ and $\lambda > 0$, the following inequality holds*:

$$\frac{1}{2\lambda}d^2(J_\lambda x, y) - \frac{1}{2\lambda}d^2(x, y) + \frac{1}{2\lambda}d^2(x, J_\lambda x) + f(J_\lambda x) \le f(y).$$

Lemma 7 [17, 27] (The resolvent identity). *Let (X, d) be a complete* CAT(0) *space and $f : X \to (-\infty, \infty]$ be proper convex and lower semi-continuous. Then the following identity holds*:

$$J_\lambda x = \Big(\frac{\lambda-\mu}{\lambda}J_\lambda x \oplus \frac{\mu}{\lambda}x\Big)$$

for all $x \in X$ and $\lambda > \mu > 0$.

Lemma 8 [11]. *If C is a closed convex subset of a complete* CAT(0) *space X and $T : C \to X$ be a uniformly L-Lipschitzian and total asymptotically nonexpansive mappings. Let $\{x_n\}$ be a bounded sequence in C such that $x_n \rightharpoonup p$ and*

$$\lim_{n\to\infty} d(x_n, Tx_n) = 0.$$

Then $Tp = p$.

Lemma 9 [25]. *Let $\{a_n\}$ be a sequence of real numbers such that there exists a subsequence $\{n_i\}$ of $\{n\}$ such that $a_{n_i} < a_{n_i+1}$ for all $i \in \mathbb{N}$. Then there exists a nondecreasing sequence $\{m_k\} \subset \mathbb{N}$ such that $m_k \to \infty$ and the following properties are satisfied by all (sufficiently large) numbers $k \in \mathbb{N}$*

$$a_{m_k} \le a_{m_k+1} \quad \textit{and} \quad a_k \le a_{m_k+1}.$$

In fact,

$$m_k = \max\{j \le k : a_j < a_{j+1}\}.$$

Lemma 10. (Xu, [30]) *Let $\{a_n\}$ be a sequence of nonnegative real numbers satisfying the following relation*:

$$a_{n+1} \le (1-\alpha_n)a_n + \alpha_n\sigma_n + \gamma_n, \quad n \ge 0,$$

where, (i) $\{\alpha_n\} \subset [0,1]$, $\sum \alpha_n = \infty$; (ii) $\limsup \sigma_n \le 0$; (iii) $\gamma_n \ge 0$; $(n \ge 0)$, $\sum \gamma_n < \infty$. Then, $a_n \to 0$ as $n \to \infty$.

3. Main Result

Theorem 1. *Let X be a complete* CAT(0) *space and C be a nonempty closed convex subset of X. Let $f : C \to (-\infty, \infty]$ be a proper convex and lower semi-continuous function and $T : C \to C$ be $L-$Lipschitzian and total asymptotically nonexpansive mappings with $\{u_n\}, \{v_n\}$ and mappings $\zeta : [0,\infty) \to [0,\infty)$ satisfying $\sum_{n=1}^{\infty} u_n < \infty$ and $\sum_{n=1}^{\infty} v_n < \infty$ such that*

$$\Omega := F(T) \bigcap \mathop{arg\min}_{y\in C} f(y) \ne \emptyset.$$

Let $\{x_n\}_{n=1}^{\infty}$ be a sequence generated by $x_1 = w \in C$ chosen arbitrarily,

$$\begin{cases} z_n = \mathop{arg\min}_{y\in C}\Big[f(y) + \dfrac{1}{2\lambda_n}d^2(y,x_n)\Big], \\ y_n = \alpha_n w \oplus (1-\alpha_n)z_n, \\ x_{n+1} = (1-\beta_n)x_n \oplus \beta_n T^n y_n, \end{cases} \tag{3.1}$$

where $\{\alpha_n\}_{n=1}^{\infty} \subset (0,1)$, $\{\beta_n\}_{n=1}^{\infty} \subset [c,d] \subset (0,1)$ satisfying

$$\lim_{n\to\infty} \alpha_n = 0, \quad \sum_{n=1}^{\infty} \alpha_n = \infty, \quad \lim_{n\to\infty} \frac{u_n}{\alpha_n} = 0, \quad \lim_{n\to\infty} \frac{v_n}{\alpha_n} = 0.$$

Assume there exists constant $M > 0$, such that $\zeta(r) \le Mr, \forall r \ge 0$. Then $\{x_n\}_{n=1}^{\infty}$ converges strongly to Ω.

P r o o f. Let $p \in \Omega$ and $f(p) \le f(y), \quad \forall\, y \in C$. Therefore we obtain

$$f(p) + \frac{1}{2\lambda_n}d^2(p,p) \le f(y) + \frac{1}{2\lambda_n}d^2(y,p), \quad \forall\, y \in C,$$

hence $p = J_{\lambda_n} p$, $\forall\ n \geq 1$. Indeed $z_n = J_{\lambda_n} x_n$ and J_{λ_n} is nonexpansive [17]. Thus

$$d(z_n, p) = d(J_{\lambda_n} x_n, J_{\lambda_n} p) \leq d(x_n, p).$$

Let $\delta_n := \alpha_n \beta_n (1 + u_n M)$. Since there exists $N_0 > 0$ such that

$$\frac{u_n}{\alpha_n} \leq \frac{\epsilon(1 + u_n M)}{M}, \quad \frac{v_n}{\alpha_n} \leq (1 + u_n M),$$

for all $n \geq N_0$ and for some $\epsilon > 0$ satisfying $0 \leq (1-\epsilon)\delta_n \leq 1$. For any point $p \in \Omega$ and $n \geq N_0$, then we have from (3.1) and from Lemma 1 that

$$\begin{aligned}
d(x_{n+1}, p) &= d((1-\beta_n)x_n \oplus \beta_n T^n y_n, p) \\
&\leq (1-\beta_n) d(x_n, p) + \beta_n d(T^n y_n, p) \\
&\leq (1-\beta_n) d(x_n, p) + \beta_n (1 + M u_n) d(y_n, p) + \beta_n v_n \\
&= (1-\beta_n) d(x_n, p) + \beta_n (1 + M u_n)[d(\alpha_n w \oplus (1-\alpha_n) z_n, p)] + \beta_n v_n \\
&\leq (1-\beta_n) d(x_n, p) + \alpha_n \beta_n (1 + M u_n) d(w, p) + \beta_n (1-\alpha_n)(1 + M u_n) d(z_n, p) + \beta_n v_n \\
&\leq [1 - \beta_n + \beta_n (1-\alpha_n)(1 + M u_n)] d(x_n, p) \\
&\quad + \alpha_n \beta_n (1 + M u_n) d(w, p) + \beta_n v_n \\
&\leq [1 - (1-\epsilon)\delta_n] d(x_n, p) + \delta_n (1-\epsilon) \frac{(d(w,p) + 1)}{(1-\epsilon)} \\
&\leq \max\left\{ d(x_n, p), \frac{(d(w,p)+1)}{(1-\epsilon)} \right\}.
\end{aligned}$$

Thus, by induction

$$d(x_n, p) \leq \max\left\{ d(x_{N_0}, p), \frac{(d(w,p)+1)}{(1-\epsilon)} \right\}, \quad \forall\ n \geq N_0.$$

It implies that $\{x_n\}$ is bounded, it follows that $\{y_n\}$ and $\{z_n\}$ are also bounded. Furthermore, from (3.1) and Lemma 2 and letting $\bar{u}_n := 2Mu_n + u_n^2$, we obtain

$$\begin{aligned}
d^2(x_{n+1}, p) &= d^2((1-\beta_n)x_n \oplus \beta_n T^n y_n, p) \\
&\leq (1-\beta_n) d^2(x_n, p) + \beta_n d^2(T^n y_n, p) - \beta_n (1-\beta_n) d^2(x_n, T^n y_n) \\
&\leq (1-\beta_n) d^2(x_n, p) + \beta_n ((1 + M u_n) d(y_n, p) + v_n)^2 - \beta_n (1-\beta_n) d^2(x_n, T^n y_n) \\
&= (1-\beta_n) d^2(x_n, p) + \beta_n (1 + M\bar{u}_n) d^2(y_n, p) + \beta_n v_n [2(1 + M u_n) d(y_n, p) + v_n] \\
&\quad - \beta_n (1-\beta_n) d^2(x_n, T^n y_n), \qquad (3.2)
\end{aligned}$$

also from Lemma 3, we have

$$\begin{aligned}
d^2(y_n, p) &= d^2(\alpha_n w \oplus (1-\alpha_n) z_n, p) \\
&\leq \alpha_n^2 d^2(w, p) + (1-\alpha_n)^2 d^2(z_n, p) + 2\alpha_n (1-\alpha_n) \langle \overrightarrow{wp}, \overrightarrow{z_n p} \rangle \\
&\leq \alpha_n^2 d^2(w, p) + (1-\alpha_n) d^2(z_n, p) + 2\alpha_n (1-\alpha_n) \langle \overrightarrow{wp}, \overrightarrow{z_n p} \rangle \qquad (3.3) \\
&\leq \alpha_n^2 d^2(w, p) + (1-\alpha_n) d^2(x_n, p) + 2\alpha_n (1-\alpha_n) \langle \overrightarrow{wp}, \overrightarrow{z_n p} \rangle. \qquad (3.4)
\end{aligned}$$

From (3.2) and (3.4) and the fact that $\{y_n\}$ is bounded, we have that there exists $D > 0$ such that for any $n \geq N_0$, $d(y_n, p) \leq D$ and letting $\theta_n := \alpha_n\beta_n$, we obtain

$$\begin{aligned}
d^2(x_{n+1}, p) &\leq (1-\beta_n)d^2(x_n, p) + \beta_n[\alpha_n^2 d^2(w, p) + (1-\alpha_n)d^2(x_n, p) + 2\alpha_n(1-\alpha_n)\langle\overrightarrow{wp}, \overrightarrow{z_n p}\rangle] \\
&\quad +\beta_n\bar{u}_n d^2(y_n, p)\beta_n v_n[2(1+Mu_n)d(y_n, p) + v_n] - \beta_n(1-\beta_n)d^2(x_n, T^n y_n) \\
&\leq (1-\theta_n)d^2(x_n, p) + \theta_n[\alpha_n d^2(w, p) + (1-\alpha_n)\langle\overrightarrow{wp}, \overrightarrow{z_n p}\rangle] \\
&\quad +\beta_n[\bar{u}_n D^2 + 2Dv_n(1+Mu_n) + v_n^2] - \beta_n(1-\beta_n)d^2(x_n, T^n y_n) \qquad (3.5) \\
&\leq (1-\theta_n)d^2(x_n, p) + \theta_n[\alpha_n d^2(w, p) + (1-\alpha_n)\langle\overrightarrow{wp}, \overrightarrow{z_n p}\rangle] \\
&\quad +\beta_n[\bar{u}_n D^2 + 2Dv_n(1+Mu_n) + v_n^2]. \qquad (3.6)
\end{aligned}$$

To complete the proof, we have to consider the following two cases.

Case 1. Suppose $\{d(x_n, p)\}$ is non-increasing, then $\{d(x_n, p)\}$ is convergent, from (3.5) and boundedness of $\{z_n\}$, then there exists $D_1 > 0$ such that for any $n \geq N_0$, $d(z_n, p) \leq D_1$, thus

$$\begin{aligned}
\beta_n(1-\beta_n)d^2(x_n, T^n y_n) &\leq d^2(x_n, p) - d^2(x_{n+1}, p) \\
&\quad +\theta_n[\alpha_n d^2(w, p) + 2(1-\alpha_n)\langle\overrightarrow{wp}, \overrightarrow{z_n p}\rangle - d^2(x_n, p)] \\
&\quad +\beta_n[\bar{u}_n D^2 + 2Dv_n(1+Mu_n) + v_n^2] \\
&\leq d^2(x_n, p) - d^2(x_{n+1}, p) \\
&\quad +\theta_n[\alpha_n d^2(w, p) + 2(1-\alpha_n)D_1 d(w, p) - d^2(x_n, p)] \\
&\quad +\beta_n[\bar{u}_n D^2 + 2Dv_n(1+Mu_n) + v_n^2], \qquad (3.7)
\end{aligned}$$

which implies that

$$\beta_n(1-\beta_n)d^2(x_n, T^n y_n) \to 0 \quad \text{as} \quad n \to \infty,$$

hence

$$\lim_{n\to\infty} d(x_n, T^n y_n) = 0 \qquad (3.8)$$

and from (3.1), we have

$$d(y_n, z_n) \leq \alpha_n d(w, z_n) + (1-\alpha_n)d(z_n, z_n) \to 0 \quad \text{as} \quad n \to \infty, \qquad (3.9)$$

also from (3.1) and (3.8), we obtain

$$d(x_{n+1}, x_n) \leq (1-\beta_n)d(x_n, x_n) + \beta_n d(T^n y_n, x_n) \to 0 \quad \text{as} \quad n \to \infty. \qquad (3.10)$$

Furthermore from Lemma 6, we see that

$$\frac{1}{2\lambda_n}d^2(z_n, p) - \frac{1}{2\lambda_n}d^2(x_n, p) + \frac{1}{2\lambda_n}d^2(x_n, z_n) \leq f(p) - f(z_n),$$

since $f(p) \leq f(z_n)$ for all $n \geq 1$, it follows that

$$d^2(x_n, z_n) \leq d^2(x_n, p) - d^2(z_n, p). \qquad (3.11)$$

But from (3.2) and (3.3), we obtain

$$\begin{aligned}
d^2(x_{n+1}, p) \leq (1-\beta_n)d^2(x_n, p) + \beta_n[\alpha_n^2 d^2(w, p) + (1-\alpha_n)d^2(z_n, p) + 2\alpha_n(1-\alpha_n)\langle\overrightarrow{wp}, \overrightarrow{z_n p}\rangle] \\
+\beta_n[2v_n(1+Mu_n)d(y_n, p) + v_n^2 + \bar{u}_n d^2(y_n, p)],
\end{aligned}$$

therefore, from (2.1) and boundedness of $\{z_n\}$, we obtain

$$\begin{aligned} d^2(x_n,p) \leq & \frac{1}{\beta_n}(d^2(x_n,p)-d^2(x_{n+1},p))+\alpha_n^2 d^2(w,p)+(1-\alpha_n)d^2(z_n,p) \\ & +2\alpha_n(1-\alpha_n)D_1 d(w,p)+2Dv_n(1+Mu_m)+\bar{u}_n D^2, \end{aligned} \tag{3.12}$$

from (3.11) and (3.12), we obtain

$$\begin{aligned} d^2(x_n,z_n) \leq & \frac{1}{\beta_n}(d^2(x_n,p)-d^2(x_{n+1},p))+\alpha_n(\alpha_n d^2(w,p)-d^2(z_n,p)) \\ & +2\alpha_n(1-\alpha_n)D_1 d(w,p)+2Dv_n(1+Mu_m)+\bar{u}_n D^2, \end{aligned}$$

since $\{x_n\}$ and $\{z_n\}$ are bounded and $\{d(x_n,p)\}$ is non-increasing sequence, it follows from that

$$\lim_{n\to\infty} d(x_n,z_n)=0, \tag{3.13}$$

from (3.9) and (3.13), we obtain

$$\lim_{n\to\infty} d(x_n,y_n)=0 \tag{3.14}$$

and

$$d(y_n,T^n y_n) \leq d(y_n,x_n)+d(x_n,T^n y_n) \to 0 \quad \text{as} \quad n\to\infty, \tag{3.15}$$

also from (3.14) and (3.15), we obtain

$$\begin{aligned} d(x_n,T^n x_n) & \leq d(x_n,y_n)+d(y_n,T^n y_n)+d(T^n y_n,T^n x_n) \\ & \leq (2+Mu_n)d(x_n,y_n)+d(y_n,T^n y_n)+v_n \to 0 \end{aligned} \tag{3.16}$$

as $n\to\infty$. Observe also that since T is uniformly L-Lipschitzian, we have

$$\begin{aligned} d(x_n,Tx_n) & \leq d(x_n,T^n x_n)+d(T^n x_n,Tx_n) \\ & \leq d(x_n,T^n x_n)+Ld(T^{n-1}x_n,x_n) \\ & \leq d(x_n,T^n x_n)+L[d(T^{n-1}x_n,T^{n-1}x_{n-1})+d(T^{n-1}x_{n-1},x_{n-1})+d(x_{n-1},x_n)] \\ & \leq d(x_n,T^n x_n)+Ld(T^{n-1}x_{n-1},x_{n-1})+L(1+L)d(x_n,x_{n-1}), \end{aligned}$$

it follows from (3.10) and (3.16) that

$$\lim_{n\to\infty} d(x_n,Tx_n)=0, \tag{3.17}$$

from (3.14) and (3.17), we obtain

$$\begin{aligned} d(y_n,Ty_n) & \leq d(y_n,x_n)+d(x_n,Tx_n)+d(Tx_n,Ty_n) \\ & \leq (1+L)d(y_n,x_n)+d(x_n,Tx_n) \to 0 \end{aligned}$$

as $n\to\infty$. Also since $\lambda_n \geq \lambda > 0$, from Lemma 7, we obtain

$$\begin{aligned} d(J_\lambda x_n,J_{\lambda_n}x_n) & = d\Big(J_\lambda x_n,J_\lambda\Big(\frac{\lambda_n-\lambda}{\lambda_n}J_{\lambda_n}x_n \oplus \frac{\lambda}{\lambda_n}x_n\Big)\Big) \\ & \leq d(x_n,(1-\frac{\lambda}{\lambda_n})J_{\lambda_n}x_n \oplus \frac{\lambda}{\lambda_n}x_n) \\ & \leq (1-\frac{\lambda}{\lambda_n})d(x_n,z_n) \to 0 \end{aligned}$$

as $n \to \infty$, hence

$$d(x_n, J_\lambda x_n) \leq d(x_n, z_n) + d(z_n, J_\lambda x_n) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.18}$$

Moreover, since $\{x_n\}$ is bounded and X is a complete CAT(0) space, by Lemma 4 we choose a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ such that $\triangle - \lim x_{n_i} = v$, where $v := P_\Omega(w)$. Then, from (3.15), (3.18), Lemma 8 and the fact that J_λ is nonexpansive [17], we have $v \in F(T)$, also from Lemma 5, we have

$$\limsup \langle \overrightarrow{wv}, \overrightarrow{x_n v} \rangle \leq 0. \tag{3.19}$$

Furthermore, since

$$\begin{aligned} \langle \overrightarrow{wv}, \overrightarrow{z_n v} \rangle &= \langle \overrightarrow{wv}, \overrightarrow{z_n x_n} \rangle + \langle \overrightarrow{wv}, \overrightarrow{x_n v} \rangle \\ &\leq d(w, v) d(z_n, x_n) + \langle \overrightarrow{wv}, \overrightarrow{x_n v} \rangle, \end{aligned}$$

it follows from (3.13) and (3.19) that

$$\limsup \langle \overrightarrow{wv}, \overrightarrow{z_n v} \rangle \leq 0.$$

Thus, now putting $v := p$ in inequality (3.6), we get that, for $n \geq N_0$

$$\begin{aligned} d^2(x_{n+1}, v) \leq (1 - \theta_n) d^2(x_n, v) + \theta_n [\alpha_n d^2(w, v) + (1 - \alpha_n) \langle \overrightarrow{wv}, \overrightarrow{z_n v} \rangle] \\ + \beta_n [\bar{u}_n D^2 + 2 D v_n (1 + M u_n) + v_n^2]. \end{aligned} \tag{3.20}$$

Hence

$$d^2(x_{n+1}, v) \leq (1 - \theta_n) d^2(x_n, v) + \theta_n \sigma_n + \gamma_n,$$

where

$$\sigma_n := \alpha_n d^2(w, v) + (1 - \alpha_n) \langle \overrightarrow{wv}, \overrightarrow{z_n v} \rangle, \quad \gamma_n := \beta_n [\bar{u}_n D^2 + 2 D v_n (1 + M u_n) + v_n^2],$$

it follows from Lemma 10 that $d(x_n, v) \to 0$ as $n \to \infty$. Consequently, $x_n \to v$.

Case 2. Suppose that $\{d(x_n, p)\}_{n \geq 1}$ is non-decreasing sequence. Then, there exists a subsequence $\{n_i\}$ of $\{n\}$ such that

$$d(x_{n_i}, p) < d(x_{n_i + 1}, p)$$

for all $i \in \mathbb{N}$. Then, by Lemma 9, there exists an increasing sequence $\{m_j\}_{j \geq 1}$ such that $m_j \to \infty$, $d(x_{m_j}, p) \leq d(x_{m_j+1}, p)$ and $d(x_j, p) \leq d(x_{m_j+1}, p)$ for all $j \geq 1$. Then from (3.7), we obtain

$$\begin{aligned} \beta_{m_j}(1 - \beta_{m_j}) d^2(x_{m_j}, T^{m_j} y_{m_j}) &\leq d^2(x_{m_j}, p) - d^2(x_{m_j+1}, p) \\ &\quad + \theta_{m_j} [\alpha_{m_j} d^2(w, p) + 2(1 - \alpha_{m_j}) \langle \overrightarrow{wp}, \overrightarrow{z_{m_j} p} \rangle - d^2(x_{m_j}, p)] \\ &\quad + \beta_{m_j} [\bar{u}_{m_j} D^2 + 2 D v_{m_j} (1 + M u_{m_j}) + v_{m_j}^2] \\ &\leq d^2(x_{m_j}, p) - d^2(x_{m_j+1}, p) \\ &\quad + \theta_{m_j} [\alpha_{m_j} d^2(w, p) + 2(1 - \alpha_{m_j}) D_1 d(w, p) - d^2(x_{m_j}, p)] \\ &\quad + \beta_{m_j} [\bar{u}_{m_j} D^2 + 2 D v_{m_j} (1 + M u_{m_j}) + v_{m_j}^2]. \end{aligned}$$

This implies $d(x_{m_j}, T^{m_j} y_{m_j}) \to 0$ as $j \to \infty$. Thus, as in Case 1, we obtain that $d(x_{m_j}, T x_{m_j}) \to 0$ and $d(x_{m_j}, J_\lambda x_{m_j}) \to 0$ as $j \to \infty$ and also following the same argument in Case 1, we get $\limsup \langle \overrightarrow{wv}, \overrightarrow{z_{m_j} v} \rangle \leq 0$, where $v := P_\Omega(w)$. Also from (3.20), we obtain that,

$$\begin{aligned} d^2(x_{m_j+1}, v) \leq (1 - \theta_n) d^2(x_{m_j}, v) + \theta_{m_j} [\alpha_{m_j} d^2(w, v) + (1 - \alpha_{m_j}) \langle \overrightarrow{wv}, \overrightarrow{z_{m_j} v} \rangle] \\ + \beta_{m_j} [\bar{u}_{m_j} D^2 + 2 D v_{m_j} (1 + M u_{m_j}) + v_{m_j}^2]. \end{aligned} \tag{3.21}$$

Since $d^2(x_{m_j}, v) \le d^2(x_{m_j+1}, v)$, it follows that

$$\begin{aligned}\theta_{m_j} d^2(x_{m_j}, v) &\le d^2(x_{m_j}, v) - d^2(x_{m_j+1}, v) + \theta_{m_j}[\alpha_{m_j} d^2(w, v) + (1-\alpha_{m_j})\langle \overrightarrow{wv}, \overrightarrow{z_{m_j}v}\rangle] \\ &\quad + \beta_{m_j}[\bar{u}_{m_j} D^2 + 2Dv_{m_j}(1+Mu_{m_j}) + v_{m_j}^2] \\ &\le \theta_{m_j}[\alpha_{m_j} d^2(w, v) + (1-\alpha_{m_j})\langle \overrightarrow{wv}, \overrightarrow{z_{m_j}v}\rangle] \\ &\quad + \beta_{m_j}[\bar{u}_{m_j} D^2 + 2Dv_{m_j}(1+Mu_{m_j}) + v_{m_j}^2].\end{aligned}$$

In particular, since $\theta_{m_j} > 0$, we get

$$d^2(x_{m_j}, v) \le [\alpha_{m_j} d^2(w, v) + (1-\alpha_{m_j})\langle \overrightarrow{wv}, \overrightarrow{z_{m_j}v}\rangle] + \Big[\frac{\bar{u}_{m_j}}{\alpha_{m_j}} D^2 + 2D\frac{v_{m_j}}{\alpha_{m_j}}(1+Mu_{m_j}) + v_{m_j}\frac{v_{m_j}}{\alpha_{m_j}}\Big].$$

Then, since $\limsup \langle \overrightarrow{wv}, \overrightarrow{x_{m_j}v}\rangle \le 0$ and the fact that $\alpha_{m_j} \to 0$ as $j \to \infty$ and

$$\lim_{j\to\infty} \frac{u_{m_j}}{\alpha_{m_j}} = 0, \quad \lim_{j\to\infty} \frac{v_{m_j}}{\alpha_{m_j}} = 0$$

we obtain that $d(x_{m_j}, v) \to 0$ as $j \to \infty$. This together with (3.21) give $d(x_{m_j+1}, v) \to 0$ as $j \to \infty$. But $d(x_j, v) \le d(x_{m_j+1}, v)$, for all $j \ge 1$, thus we obtain that $x_j \to v$. Therefore, from the above two cases, we can conclude that $\{x_n\}_{n=1}^{\infty}$ converges strongly to an element of Ω and the proof is complete. □

4. Conclusion

In this work, we study a modified Halpern-type proximal point algorithm for finding the minimizer of a convex lower semi-continuous function which is also a fixed point of total asymptotically nonexpansive mapping. Under some appropriate assumption, we have obtained a strong convergence theorem for the proposed algorithm in the framework of a complete CAT(0) space.

REFERENCES

1. Alber Ya. I., Chidume C. E., Zegeye H. Approximating fixed points of total asymptotically nonexpansive mappings. *Fixed Point Theory Appl.*, 2006. Art. no. 10673. P. 1–20. DOI: 10.1155/FPTA/2006/10673
2. Ahmad I., Ahmad M. An implicit viscosity technique of nonexpansive mapping in CAT(0) spaces. *Open J. Math. Anal.*, 2017. Vol. 1. P. 1–12. DOI: 10.30538/psrp-oma2017.0001
3. Agarwal R. P., O'Regan D., Sahu D. R. Iterative construction of fixed points of nearly asymptotically nonexpansive mappings. *J. Nonlinear Convex Anal.*, 2007. Vol. 8, No. 1. P. 61–79.
4. Ambrosio L., Gigli N., Savare G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed. Lectures in Mathematics ETH Zürich. Basel: Birkhäuser, 2008. 334 p. DOI: 10.1007/978-3-7643-8722-8
5. Ariza-Ruiz D., Leuştean L., López-Acedo G. Firmly nonexpansive mappings in classes of geodesic spaces. *Trans. Amer. Math. Soc.*, 2014. Vol. 366. No. 8. P. 4299–4322. DOI: 10.1090/S0002-9947-2014-05968-0
6. Bačák M. The proximal point algorithm in metric spaces. *Israel J. Math.*, 2013. Vol. 194. P. 689–701. DOI: 10.1007/s11856-012-0091-3
7. Berg I. D., Nikolaev I. G. Quasilinearization and curvature of Aleksandrov spaces. *Geom. Dedicata*, 2008. Vol. 133. P. 195–218. DOI: 10.1007/s10711-008-9243-3
8. Bonyah E., Ahmad M., Ahmad I. On the viscosity rule for common fixed points of two nonexpansive mappings in CAT(0) spaces. *Open J. Math. Sci.*, 2018. Vol. 2. No. 1. P. 39–55. DOI: 10.30538/oms2018.0016
9. Bridson M. R., Häfliger A. *Metric Spaces of Nonpositive Curvature.* Grundlehren Math. Wiss., vol. 319. Berlin, Heidelberg: Springer-Verlag, 1999. 643 p. DOI: 10.1007/978-3-662-12494-9
10. Burago D., Burago Yu., Ivanov S. *A Course in Metric Geometry.* Grad. Stud. Math., vol. 33. Providence, RI: A.M.S., 2001. 415 p.

11. Chang S.-S., Wang L., Joseph Lee H. W., Chan C. K., Yang L. Demiclosed principle and △-convergence theorems for total asymptotically nonexpansive mappings in CAT(0) spaces. *Appl. Math. Comput.*, 2012. Vol. 219, No. 5. P. 2611–2617. DOI: 10.1016/j.amc.2012.08.095
12. Chang S.-S., Yao J.-C., Wang L., Qin L. J. Some convergence theorems involving proximal point and common fixed points for asymptotically nonexpansive mappings in CAT(0) spaces. *Fixed Point Theory Appl.*, 2016. Art. no. 68. P. 1–11. DOI: 10.1186/s13663-016-0559-7
13. Cholamjiak P., Abdou A. A., Cho Y. J. Proximal point algorithms involving fixed points of nonexpansive mappings in CAT(0) spaces. *Fixed Point Theory Appl.*, 2015. Art. no. 227. P. 1–13. DOI: 10.1186/s13663-015-0465-4
14. Dehghan H., Rooin J. *A Characterization of Metric Projection in* CAT(0) *spaces*. 2012. 3 p. arXiv: 1311.4174 [math.FA]
15. Dhompongsa S., Kirk W. A., Sims B. Fixed points of uniformly lipschitzian mappings. *Nonlinear Anal.*, 2006. Vol. 65, No. 4. P. 762–772. DOI: 10.1016/j.na.2005.09.044
16. Dhompongsa S., Panyanak B. On △-convergence theorems in CAT(0) spaces. *Comput. Math. Appl.*, 2008. Vol. 56, No. 10. P. 2572–2579. DOI: 10.1016/j.camwa.2008.05.036
17. Jost J. Convex functionals and generalized harmonic maps into spaces of non positive curvature. *Comment. Math. Helv.*, 1995. Vol. 70. P. 659–673. DOI: 10.1007/BF02566027
18. Güler O. On the convergence of the proximal point algorithm for convex minimization. *SIAM J. Control Optim.*, 1991. Vol. 29, No. 2. P. 403–419. DOI: 10.1137/0329022
19. Kakavandi B. A. Weak topologies in complete CAT(0) metric spaces. *Proc. Amer. Math. Soc.*, 2012. Vol. 141, No. 3. P. 1029–1039. URL: https://www.jstor.org/stable/23558440
20. Kamimura S., Takahashi W. Approximating solutions of maximal monotone operators in Hilbert spaces. *J. Approx. Theory*, 2000. Vol. 106, No. 2. P. 226–240. DOI: 10.1006/jath.2000.3493
21. Kang S. M., Haq A. U., Nazeer W., Ahmad I., Ahmad M. Explicit viscosity rule of nonexpansive mappings in CAT(0) spaces. *J. Comput. Anal. Appl.*, 2019. Vol. 27, No. 6. P. 1034–1043.
22. Kirk W. A. Geodesic geometry and fixed point theory. In: *Seminar of Mathematical Analysis, Malaga/Seville, 2002/2003, Álvares D.G., Acelo G.L., Caro R.V. (eds.)*, vol. 64. P. 195–225.
23. Kirk W. A.Geodesic geometry and fixed point theory, II. In: *Int. Conf. on Fixed Point Theory and Applications, Yokohama Publ., Yokohama, Japan, 2004*. P. 113–142.
24. Kirk W. A., Panyanak B. A concept of convergence in geodesic spaces. *Nonlinear Anal.*, 2008. Vol. 68, No. 12. P. 3689–3696. DOI: 10.1016/j.na.2007.04.011
25. Maingé P. E. Strong convergence of projected subgradient methods for nonsmooth and nonstrictly convex minimization. *Set-Valued Anal.*, 2008. Vol. 16, No. 7–8. P. 899–912. DOI: 10.1007/s11228-008-0102-z
26. Martinet B. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Fr. Inform. Rech. Opér.*, 1970. Vol. 4, No. R3. P. 154–158. (in France)
27. Mayer U. F. Gradient flows on nonpositively curved metric spaces and harmonic maps. *Commun. Anal. Geom.* 1998. Vol. 6, No. 2. P. 199–253.
28. Rockafellar R. T. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 1976. Vol. 14, No. 5. P. 877–898. DOI: /10.1137/0314056
29. Suparatulatorn R., Cholamjiak P., Suantai S. On solving the minimization problem and the fixed-point problem for nonexpansive mappings in CAT(0) spaces. *Optim. Methods Softw.*, 2017. Vol. 32, No. 1. P. 182–192. DOI: 10.1080/10556788.2016.1219908
30. Xu H. K. Iterative algorithms for nonlinear operators. *J. Lond. Math. Soc. (2)*, 2002. Vol. 66, No. 1. P. 240–256. DOI: 10.1112/S0024610702003332

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 120–159
DOI: 10.15826/umj.2021.1.011

CONTROL SYSTEM DEPENDING ON A PARAMETER[1]

Vladimir N. Ushakov†, Aleksandr A. Ershov††, Andrey V. Ushakov†††, Oleg A. Kuvshinov††††

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, 620108, Russia

†ushak@imm.uran.ru, ††ale10919@yandex.ru, †††aushakov.pk@gmail.com,
††††okuvshinov@inbox.ru

Abstract: A nonlinear control system depending on a parameter is considered in a finite-dimensional Euclidean space and on a finite time interval. The dependence on the parameter of the reachable sets and integral funnels of the corresponding differential inclusion system is studied. Under certain conditions on the control system, the degree of this dependence on the parameter is estimated. Problems of targeting integral funnels to a target set in the presence of an obstacle in strict and soft settings are considered. An algorithm for the numerical solution of this problem in the soft setting has been developed. An estimate of the error of the developed algorithm is obtained. An example of solving a specific problem for a control system in a two-dimensional phase space is given.

Keywords: Control system, Differential inclusion, Reachable set, Integral funnel, Parameter dependence, Approximation.

Introduction

A nonlinear control system depending on a parameter is considered in a finite-dimensional Euclidean space and on a finite time interval.

The reachable sets and integral funnels of the differential inclusion corresponding to the system are studied. The problems related to the study of reachable sets and integral funnels of dynamical systems are closely intertwined with numerous problems in the theory of dynamical systems including those that arise in control theory and the theory of differential games [5, 6, 10–13, 16, 17]. Various theoretical approaches and associated computational methods [1–3, 5–14, 16–21] are used in the study of reachable sets, their construction, and estimation. These control problems and differential games include, for example, various types of approach problems, resolving constructions of which include one of the main components that are called solvability sets, i.e., the sets of those positions of the control system from which the approach problem is solvable [10–13]. For many problems, these sets can be described quite simply in terms of reachable sets and integral funnels [1, 2, 5–9, 12, 13, 16–21]. Some problems can be formulated as problems of the theory of controllability of dynamical systems [19].

In this paper, we study the dependence on a parameter of reachable sets and integral funnels: the degree of this dependence on the parameter is estimated under certain conditions imposed on the control system. We introduce systems of sets in the phase space that approximate reachable sets and integral funnels on a given time interval corresponding to a finite partition of this interval. In this case, the degree of dependence on the parameter of the approximating system of sets is first estimated, and then this estimate is used to estimate the dependence on the parameter of

[1]This research was supported by the Russian Science Foundation (project no. 19-11-00105).

the reachable sets and integral funnels of the differential inclusion. This approach is natural and especially useful for studying specific applied control problems, when, in the end, one has to deal not with ideal reachable sets and integral funnels, but with their approximations corresponding to a discrete representation of the time interval.

1. Estimates of reachable sets and integral funnels of differential inclusions

Consider a control system Σ

$$\frac{dx}{dt} = f_\alpha(t, x, u), \quad u \in P \in \operatorname{comp}(\mathbb{R}^p) \tag{1.1}$$

on a time interval $[t_0, \vartheta]$, $t_0 < \vartheta < \infty$; here $x \in \mathbb{R}^n$ is the phase vector of Σ, u is the control vector, α is a parameter from a set $\mathscr{L} \in \operatorname{comp}(\mathbb{R}^l)$; $\operatorname{comp}(\mathbb{R}^k)$ is the set of compact subsets of $\mathbb{R}^k$ with the Hausdorff metric

$$d(X^{(1)}, X^{(2)}) = \max(h(X^{(1)}, X^{(2)}), \quad h(X^{(2)}, X^{(1)})), \quad h(X^{(1)}, X^{(2)}) = \max_{x^{(1)} \in X^{(1)}} \rho(x^{(1)}, X^{(2)})$$

is the Hausdorff deviation of $X^{(1)}$ from $X^{(2)}$, where

$$\rho(x^{(1)}, X^{(2)}) = \min_{x^{(2)} \in X^{(2)}} \|x^{(1)} - x^{(2)}\|.$$

We assume that the system Σ satisfies the following conditions.

A. The function $f_\alpha(t, x, u)$ is defined on $[t_0, \vartheta] \times \mathbb{R}^n \times P \times \mathscr{L}$ and, for any bounded and closed domain $D \subset [t_0, \vartheta] \times \mathbb{R}^n$, there are a continuous function $\omega^*(r)$, $r \in (0, \infty)$ ($\omega^*(r) \downarrow 0$, $r \downarrow 0$) and a continuous function $L(t) \in (0, \infty)$, $t \in [t_0, \vartheta]$, satisfying the relations

$$\begin{gathered}
\|f_\alpha(t, x, u) - f_\beta(\tau, x, u)\| \leqslant \omega^*(|t - \tau| + \|\alpha - \beta\|), \\
(t, x) \in D, \quad (\tau, x) \in D, \quad u \in P, \quad \alpha, \beta \in \mathscr{L}; \\
\|f_\alpha(t, x, u) - f_\alpha(t, y, u)\| \leqslant L(t)\|x - y\|, \\
(t, x) \in D, \quad (t, y) \in D, \quad u \in P, \quad \alpha \in \mathscr{L}.
\end{gathered}$$

B. There is $\gamma \in (0, \infty)$ such that

$$\|f_\alpha(t, x, u)\| \leqslant \gamma(1 + \|x\|), \quad (t, x, u) \in [t_0, \vartheta] \times \mathbb{R}^n \times P, \quad \alpha \in \mathscr{L}.$$

We introduce a multivalued mapping

$$\begin{gathered}
(t, x) \mapsto F_\alpha(t, x) = \operatorname{co} \mathcal{F}_\alpha(t, x), \\
\mathcal{F}_\alpha(t, x) = \{f_\alpha(t, x, u) : u \in P\} \in \operatorname{comp}(\mathbb{R}^n), \\
(t, x) \in [t_0, \vartheta] \times \mathbb{R}^n, \quad \alpha \in \mathscr{L}.
\end{gathered}$$

The mapping $(t, x) \mapsto F_\alpha(t, x) \in \operatorname{comp}(\mathbb{R}^n)$ satisfies the following conditions.

A*. For any bounded and closed domain $D \subset [t_0, \vartheta] \times \mathbb{R}^n$, there are a continuous function $\omega^*(r)$, $r \in (0, \infty)$ ($\omega^*(r) \downarrow 0$, $r \downarrow 0$) and a continuous function $L(t) \in (0, \infty)$, $t \in [t_0, \vartheta]$, satisfying the relations

$$\begin{gathered}
d(F_\alpha(t, x), F_\beta(\tau, x)) \leqslant \omega^*(|t - \tau| + \|\alpha - \beta\|), \\
(t, x) \in D, \quad (\tau, x) \in D, \quad \alpha, \beta \in \mathscr{L};
\end{gathered} \tag{1.2}$$

$$\begin{gathered}
d(F_\alpha(t, x), F_\alpha(t, y)) \leqslant L(t)\|x - y\|, \\
(t, x) \in D, \quad (t, y) \in D, \quad \alpha \in \mathscr{L}.
\end{gathered} \tag{1.3}$$

B*. There is $\gamma \in (0,\infty)$ such that

$$h(F_\alpha(t,x),\{\mathbf{0}\}) \leqslant \gamma \cdot (1+\|x\|), \quad (t,x,\alpha) \in [t_0,\vartheta] \times \mathbb{R}^n \times \mathscr{L};$$

here $\mathbf{0}$ is the null-vector in $\mathbb{R}^n$.

Let us introduce on $[t_0,\vartheta]$ the differential inclusion

$$\frac{dx}{dt} \in F_\alpha(t,x), \quad \alpha \in \mathscr{L}, \tag{1.4}$$

that satisfies the system Σ.

Let t_* and t^* ($t_* < t^*$) be from $[t_0,\vartheta]$, $x_* \in \mathbb{R}^n$, $X_* \in \mathrm{comp}\,(\mathbb{R}^n)$, and $\alpha \in \mathscr{L}$.

Let us introduce the notation:

- $X_\alpha(t^*,t_*,x_*)$ is the reachable set of the differential inclusion (1.4) at the time t^* with the initial point $x(t_*) = x_*$;
- $X_\alpha(t^*,t_*,X_*) = \bigcup\limits_{x_* \in X_*} X_\alpha(t^*,t_*,x_*)$ is the reachable set of the differential inclusion (1.4) at the time t^* with the initial set X_*.

It is known that $X_\alpha(t^*,t_*,X_*) \in \mathrm{comp}\,(\mathbb{R}^n)$, the mapping $(t^*,t_*,X_*) \mapsto X_\alpha(t^*,t_*,X_*)$ is continuous in t^* on $[t_*,\vartheta]$ for fixed $(t_*,X_*) \in [t_0,\vartheta] \times \mathrm{comp}\,(\mathbb{R}^n)$ in the Hausdorff metric, and also $X_\alpha(t^*,t_*,X_*)$ continuously depends on X_* for fixed t_*, t^*, and α.

The mapping $\alpha \mapsto X_\alpha(t^*,t_*,X_*)$ is also continuous on $\mathscr{L}$ for fixed (t^*,t_*,X_*), $t_0 \leqslant t_* < t^* \leqslant \vartheta$, and $X_* \in \mathrm{comp}\,(\mathbb{R}^n)$.

Let us refine the continuous dependence of $\alpha \mapsto X_\alpha(t^*,t_*,X_*)$ on the set $\mathscr{L}$. To do this, we derive an upper bound for the Hausdorff distance

$$d(X_\alpha(t^*,t_*,X_*), X_\beta(t^*,t_*,X_*)), \quad \alpha,\beta \in \mathscr{L}, \tag{1.5}$$

which we represent as a function of $\|\alpha-\beta\|$.

It is known that, under the conditions **A*** and **B***, the reachable set $X_\alpha(t^*,t_*,X_*)$ satisfies the equality

$$X_\alpha(t^*,t_*,X_*) = \lim_{\Delta=\Delta(\Gamma)\downarrow 0} \widetilde{X}_\alpha^{\Gamma_*}(t^*).$$

Here $\widetilde{X}_\alpha^{\Gamma_*}(t^*) \subset \mathbb{R}^n$, $\alpha \in \mathscr{L}$ are the sets corresponding to the partition

$$\Gamma_* = \{\tau_0 = t_*, \tau_1, ..., \tau_i, ..., \tau_N = t^*\}$$
$$(\tau_{i+1} - \tau_i = \Delta = \Delta(\Gamma_*) = N^{-1}(t^*-t_*), \quad i = \overline{0,N-1})$$

of the interval $[t_*,t^*]$ defined by the equality $\widetilde{X}_\alpha^{\Gamma_*}(t^*) = \widetilde{X}_\alpha^{\Gamma_*}(t_N)$ and the recurrence relations

$$\widetilde{X}_\alpha^{\Gamma_*}(\tau_0) = X_*, \quad \widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}) = \widetilde{X}_\alpha(\tau_{i+1},\tau_i,\widetilde{X}_\alpha^{\Gamma_*}(\tau_i)), \quad i = \overline{0,N-1}, \tag{1.6}$$

where

$$\widetilde{X}_\alpha(\tau^*,\tau_*,W_*) = \big\{x^* \in \mathbb{R}^n : x^* = w_* + (\tau^*-\tau_*)f_*,\ w_* \in W_*,\ f_* \in F_\alpha(\tau_*,w_*)\big\},$$
$$t_* \leqslant \tau_* < \tau^* \leqslant t^*, \quad W_* \in \mathrm{comp}\,(\mathbb{R}^n).$$

Taking into account the condition **B*** and the size of the compact set X_*, we can specify a bounded and closed domain $D \subset [t_0,\vartheta] \times \mathbb{R}^n$ containing all sets arising in the subsequent reasoning and estimates in the space $[t_0, \vartheta] \times \mathbb{R}^n$. We assume that functions $\omega^*(r)$, $r \in (0, \infty)$, and $L(t)$, $t \in [t_0, \vartheta]$, corresponding to this domain D are used in further estimates.

We first estimate quantity (1.5) for a one-point set $X_* = \{x_*\}$, $(t_*, x_*) \in D$.

When deriving an estimate for quantity (1.6), we will apply the so-called "step-by-step" reasoning scheme and "step-by-step" estimates, that is, we will move through the steps $[\tau_i, \tau_{i+1}]$, $i = \overline{0, N-1}$, of the partition Γ_*.

We start deriving an estimate with the interval $[\tau_0, \tau_1]$ of the partition Γ_*. Let us find an upper bound for the Hausdorff deviation

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_1), \widetilde{X}_\beta^{\Gamma_*}(\tau_1)), \quad \alpha, \beta \in \mathscr{L};$$

here $\widetilde{X}_\alpha^{\Gamma_*}(\tau_1) = \widetilde{X}_\alpha(\tau_1, \tau_0, x_*)$ and $\widetilde{X}_\beta^{\Gamma_*}(\tau_1) = \widetilde{X}_\beta(\tau_1, \tau_0, x_*)$.

In $\widetilde{X}_\alpha^{\Gamma_*}(\tau_1)$, we choose a point $x(\tau_1)$ such that $\rho(x(\tau_1), \widetilde{X}_\beta^{\Gamma_*}(\tau_1)) = h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_1), \widetilde{X}_\beta^{\Gamma_*}(\tau_1))$. The point $x(\tau_1)$ is representable as

$$x(\tau_1) = x_* + \Delta f_\alpha(\tau_0), \quad f_\alpha(\tau_0) \in F_\alpha(\tau_0, x_*).$$

Let us choose a vector $f_\beta(\tau_0)$ in $F_\beta(\tau_0, x_*)$ closest to $f_\alpha(\tau_0)$. The following estimate is valid:

$$\|f_\alpha(\tau_0) - f_\beta(\tau_0)\| = \rho(f_\alpha(\tau_0), F_\beta(\tau_0, x_*)) \leqslant h(F_\alpha(\tau_0, x_*), F_\beta(\tau_0, x_*)) \leqslant \omega^*(\|\alpha - \beta\|).$$

In $\widetilde{X}_\beta^{\Gamma_*}(\tau_1)$, we consider the point $y(\tau_1) = x_* + \Delta f_\beta(\tau_0)$, $\Delta = \Delta(\Gamma_*)$. There is an estimate

$$\|x(\tau_1) - y(\tau_1)\| \leqslant \Delta\omega^*(\|\alpha - \beta\|).$$

The definition of the point $x(\tau_1)$ and the inclusion $y(\tau_1) \in \widetilde{X}_\beta^{\Gamma_*}(\tau_1)$ imply the estimate

$$h(\tau_1) \leqslant \Delta\omega^*(\|\alpha - \beta\|); \tag{1.7}$$

here $h(\tau_1) = h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_1), \widetilde{X}_\beta^{\Gamma_*}(\tau_1))$.

Let us turn to the next interval $[\tau_1, \tau_2]$ of the partition Γ_* and consider the sets $\widetilde{X}_\alpha^{\Gamma_*}(\tau_2) = \widetilde{X}_\alpha(\tau_2, \tau_1, \widetilde{X}_\alpha^{\Gamma_*}(\tau_1))$ and $\widetilde{X}_\beta^{\Gamma_*}(\tau_2) = \widetilde{X}_\beta(\tau_2, \tau_1, \widetilde{X}_\beta^{\Gamma_*}(\tau_1))$.

In $\widetilde{X}_\alpha^{\Gamma_*}(\tau_2)$, we choose a point $x(\tau_2)$ such that

$$\rho(x(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)) = h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)). \tag{1.8}$$

The point $x(\tau_2)$ is representable as

$$x(\tau_2) = x_*(\tau_1) + \Delta f_\alpha(\tau_1), \quad x_*(\tau_1) \in \widetilde{X}_\alpha^{\Gamma_*}(\tau_1), \quad f_\alpha(\tau_1) \in F_\alpha(\tau_1, x_*(\tau_1)).$$

Let us choose a point $y_*(\tau_1)$ in $\widetilde{X}_\beta^{\Gamma_*}(\tau_1)$ closest to $x_*(\tau_1)$:

$$\|x_*(\tau_1) - y_*(\tau_1)\| = \rho(x_*(\tau_1), \widetilde{X}_\beta^{\Gamma_*}(\tau_1)).$$

The following estimate is valid:

$$\|x_*(\tau_1) - y_*(\tau_1)\| \leqslant h(\tau_1).$$

Let us choose a vector $f_\beta(\tau_1)$ in $F_\beta(\tau_1, y_*(\tau_1))$ closest to $f_\alpha(\tau_1)$. By (1.2) and (1.3), the following inequalities hold:

$$\begin{aligned}\|f_\alpha(\tau_1) - f_\beta(\tau_1)\| &\leqslant h(F_\alpha(\tau_1, x_*(\tau_1)), F_\beta(\tau_1, y_*(\tau_1))) \\ &\leqslant d(F_\alpha(\tau_1, x_*(\tau_1)), F_\beta(\tau_1, y_*(\tau_1))) \leqslant \omega^*(\|\alpha - \beta\|) + L(\tau_1)h(\tau_1).\end{aligned}$$

We introduce the point

$$y(\tau_2) = y_*(\tau_1) + \Delta f_\beta(\tau_1), \quad y_*(\tau_1) \in \widetilde{X}_\beta^{\Gamma_*}(\tau_1), \quad f_\beta(\tau_1) \in F_\beta(\tau_1, y_*(\tau_1)).$$

The points $x(\tau_2)$ and $y(\tau_2)$ satisfy the inequalities

$$\begin{aligned} \|x(\tau_2) - y(\tau_2)\| &\leqslant \|x_*(\tau_1) - y_*(\tau_1)\| + \Delta\|f_\alpha(\tau_1) - f_\beta(\tau_1)\| \\ &\leqslant h(\tau_1) + \Delta \cdot (\omega^*(\|\alpha - \beta\|) + L(\tau_1)h(\tau_1)) \\ &\leqslant \Delta\omega^*(\|\alpha - \beta\|) + e^{L(\tau_1)\Delta_1} \cdot h(\tau_1), \end{aligned} \tag{1.9}$$

where $\Delta_1 = \Delta = \Delta(\Gamma_*)$.

Considering (1.8) and the inclusion $y(\tau_2) \in \widetilde{X}_\beta^{\Gamma_*}(\tau_2) = \widetilde{X}_\beta(\tau_2, \tau_1, \widetilde{X}_\beta^{\Gamma_*}(\tau_1))$, we obtain

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)) \leqslant \|x(\tau_2) - y(\tau_2)\|. \tag{1.10}$$

Estimates (1.9) and (1.10) imply that

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)) \leqslant \Delta\omega^*(\|\alpha - \beta\|) + e^{L(\tau_1)\Delta_1} h(\tau_1). \tag{1.11}$$

Consider the next interval $[\tau_2, \tau_3]$ of the partition Γ_* and the sets $\widetilde{X}_\alpha^{\Gamma_*}(\tau_3) = \widetilde{X}_\alpha(\tau_3, \tau_2, \widetilde{X}_\alpha^{\Gamma_*}(\tau_2))$ and $\widetilde{X}_\beta^{\Gamma_*}(\tau_3) = \widetilde{X}_\beta(\tau_3, \tau_2, \widetilde{X}_\beta^{\Gamma_*}(\tau_2))$.

Let us find an upper bound for the Hausdorff deviation

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)), \quad \alpha, \beta \in \mathscr{L}.$$

To do this, we choose a point $x(\tau_3)$ in the $\widetilde{X}_\alpha^{\Gamma_*}(\tau_3)$ such that

$$\rho(x(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)) = h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)). \tag{1.12}$$

The point $x(\tau_3)$ is representable as

$$x(\tau_3) = x_*(\tau_2) + \Delta f_\alpha(\tau_2), \quad x_*(\tau_2) \in \widetilde{X}_\alpha^{\Gamma_*}(\tau_2), \quad f_\alpha(\tau_2) \in F_\alpha(\tau_2, x_*(\tau_2)).$$

Let us choose a point $y_*(\tau_2)$ in $\widetilde{X}_\beta^{\Gamma_*}(\tau_2)$ closest to the point $x_*(\tau_2)$:

$$\|x_*(\tau_2) - y_*(\tau_2)\| = \rho(x_*(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)).$$

The following inequality is valid

$$\|x_*(\tau_2) - y_*(\tau_2)\| \leqslant h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)).$$

Let us choose a vector $f_\beta(\tau_2)$ in $F_\beta(\tau_2, y_*(\tau_2))$ closest to $f_\alpha(\tau_2)$. We obtain the estimate

$$\begin{aligned} \|f_\alpha(\tau_2) - f_\beta(\tau_2)\| &\leqslant h(F_\alpha(\tau_2, x_*(\tau_2)), F_\beta(\tau_2, y_*(\tau_2))) \\ &\leqslant d(F_\alpha(\tau_2, x_*(\tau_2)), F_\beta(\tau_2, y_*(\tau_2))) \\ &\leqslant \omega^*(\|\alpha - \beta\|) + L(\tau_2)\|x_*(\tau_2) - y_*(\tau_2)\|. \end{aligned}$$

Consider the point $y(\tau_3) = y_*(\tau_2) + \Delta f_\beta(\tau_2)$ in $\widetilde{X}_\beta^{\Gamma_*}(\tau_3)$. The points $x(\tau_3)$ and $y(\tau_3)$ satisfy the inequalities

$$\begin{aligned} \|x(\tau_3) - y(\tau_3)\| &\leqslant \|x_*(\tau_2) - y_*(\tau_2)\| + \Delta \cdot (\omega^*(\|\alpha - \beta\|) + L(\tau_2)\|x_*(\tau_2) - y_*(\tau_2)\|) \\ &\leqslant \Delta \cdot \omega^*(\|\alpha - \beta\|) + e^{L(\tau_2)\Delta_2} h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_2), \widetilde{X}_\beta^{\Gamma_*}(\tau_2)) \\ &\leqslant \Delta \cdot \omega^*(\|\alpha - \beta\|) + e^{L(\tau_2)\Delta_2}(\Delta \cdot \omega^*(\|\alpha - \beta\|) + e^{L(\tau_1)\Delta_1} h(\tau_1)), \end{aligned}$$

where $\Delta_2 = \Delta = \Delta(\Gamma_*)$.

As a result, we get

$$\|x(\tau_3) - y(\tau_3)\| \leqslant (1 + e^{L(\tau_2)\Delta_2}) \cdot \Delta\omega^*(\|\alpha - \beta\|) + e^{L(\tau_1)\Delta_1 + L(\tau_2)\Delta_2} \cdot h(\tau_1). \tag{1.13}$$

Considering (1.12) and the inclusion $y(\tau_3) \in \widetilde{X}_\beta^{\Gamma_*}(\tau_3)$, we get

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)) \leqslant \|x(\tau_3) - y(\tau_3)\|. \tag{1.14}$$

From (1.13) and (1.14), it follows that

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)) \leqslant (1 + e^{L(\tau_2)\Delta_2}) \cdot \Delta\omega^*(\|\alpha - \beta\|) + e^{L(\tau_1)\Delta_1 + L(\tau_2)\Delta_2} \cdot h(\tau_1). \tag{1.15}$$

For a final understanding of the structure of the estimate of the quantity $h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1}))$, $i = \overline{0, N-1}$, we consider the next interval $[\tau_3, \tau_4]$ of the partition Γ_* and the sets

$$\widetilde{X}_\alpha^{\Gamma_*}(\tau_4) = \widetilde{X}_\alpha(\tau_4, \tau_3, \widetilde{X}_\alpha^{\Gamma_*}(\tau_3)) \quad \text{and} \quad \widetilde{X}_\beta^{\Gamma_*}(\tau_4) = \widetilde{X}_\beta(\tau_4, \tau_3, \widetilde{X}_\beta^{\Gamma_*}(\tau_3)).$$

Let us estimate from above the quantity

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_4), \widetilde{X}_\beta^{\Gamma_*}(\tau_4)), \quad \alpha, \beta \in \mathscr{L}.$$

To do this, we choose a point $x(\tau_4)$ in $\widetilde{X}_\alpha^{\Gamma_*}(\tau_4)$ such that

$$\rho(x(\tau_4), \widetilde{X}_\beta^{\Gamma_*}(\tau_4)) = h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_4), \widetilde{X}_\beta^{\Gamma_*}(\tau_4)).$$

The point $x(\tau_4)$ is representable in the form

$$x(\tau_4) = x_*(\tau_3) + \Delta f_\alpha(\tau_3), \quad x_*(\tau_3) \in \widetilde{X}_\alpha^{\Gamma_*}(\tau_3), \quad f_\alpha(\tau_3) \in F_\alpha(\tau_3, x_*(\tau_3)).$$

Let us choose a point $y_*(\tau_3)$ in $\widetilde{X}_\beta^{\Gamma_*}(\tau_3)$ closest to $x_*(\tau_3)$:

$$\|x_*(\tau_3) - y_*(\tau_3)\| = \rho(x_*(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)).$$

The following inequality holds:

$$\|x_*(\tau_3) - y_*(\tau_3)\| \leqslant h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_3), \widetilde{X}_\beta^{\Gamma_*}(\tau_3)). \tag{1.16}$$

Let us choose a vector $f_\beta(\tau_3)$ in $F_\beta(\tau_3, y_*(\tau_3))$ closest to $f_\alpha(\tau_3)$.

By relations (1.2) and (1.3), the following estimate is valid:

$$\begin{aligned} \|f_\alpha(\tau_3) - f_\beta(\tau_3)\| \leqslant h(F_\alpha(\tau_3, x_*(\tau_3)), F_\beta(\tau_3, y_*(\tau_3))) \leqslant d(F_\alpha(\tau_3, x_*(\tau_3)), F_\beta(\tau_3, y_*(\tau_3))) \\ \leqslant \omega^*(\|\alpha - \beta\|) + L(\tau_3)\|x_*(\tau_3) - y_*(\tau_3)\|. \end{aligned}$$

Let us choose the point $y(\tau_4) = y_*(\tau_3) + \Delta f_\beta(\tau_3)$ in $\widetilde{X}_\beta^{\Gamma_*}(\tau_4)$.

Taking into account (1.15) and (1.16), we obtain

$$\begin{gathered} \|x(\tau_4) - y(\tau_4)\| \leqslant \|x_*(\tau_3) - y_*(\tau_3)\| + \Delta\|f_\alpha(\tau_3) - f_\beta(\tau_3)\| \\ \leqslant \|x_*(\tau_3) - y_*(\tau_3)\| + \Delta\omega^*(\|\alpha - \beta\|) + L(\tau_3)\|x_*(\tau_3) - y_*(\tau_3)\| \leqslant \\ \leqslant \Delta\omega^*(\|\alpha - \beta\|) + e^{L(\tau_3)\Delta_3}\big((1 + e^{L(\tau_2)\Delta_2}) \cdot \Delta\omega^*(\|\alpha - \beta\|) + e^{L(\tau_1)\Delta_1 + L(\tau_2)\Delta_2} \cdot h(\tau_1)\big), \\ \Delta_3 = \Delta = \Delta(\Gamma_*). \end{gathered}$$

As a result, we get the estimate

$$\|x(\tau_4)-y(\tau_4)\| \leqslant \Delta\omega^*(\|\alpha-\beta\|)\cdot\big(1+e^{L(\tau_3)\Delta_3}+e^{L(\tau_3)\Delta_3+L(\tau_2)\Delta_2}\big) \\ +e^{L(\tau_3)\Delta_3+L(\tau_2)\Delta_2+L(\tau_1)\Delta_1}\cdot h(\tau_1).$$

Further, taking into account the choice of the points $x(\tau_4)$ and $y(\tau_4)$, we obtain

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_4),\widetilde{X}_\beta^{\Gamma_*}(\tau_4)) \leqslant \|x(\tau_4)-y(\tau_4)\|.$$

The latter two inequalities imply the estimate

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_4),\widetilde{X}_\beta^{\Gamma_*}(\tau_4)) \leqslant \big(1+e^{L(\tau_3)\Delta_3}+e^{L(\tau_3)\Delta_3+L(\tau_2)\Delta_2}\big)\cdot\Delta\omega^*(\|\alpha-\beta\|)+ \\ +e^{L(\tau_3)\Delta_3+L(\tau_2)\Delta_2+L(\tau_1)\Delta_1}\cdot h(\tau_1). \tag{1.17}$$

Analyzing estimates (1.11), (1.15), and (1.17), we conclude that the interval $[\tau_i,\tau_{i+1}]$, $i=\overline{1,N-1}$, of the partition Γ_* corresponds to the following estimate of the Hausdorff deviation $h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}),\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1}))$ of the set $\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}) = \widetilde{X}_\alpha(\tau_{i+1},\tau_i,\widetilde{X}_\alpha^{\Gamma_*}(\tau_i))$ from the set $\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})=\widetilde{X}_\beta(\tau_{i+1},\tau_i,\widetilde{X}_\beta^{\Gamma_*}(\tau_i))$:

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}),\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant \big(1+e^{\sum_{k=i}^{i}L(\tau_k)\Delta_k}+e^{\sum_{k=i-1}^{i}L(\tau_k)\Delta_k}+ \\ +e^{\sum_{k=i-2}^{i}L(\tau_k)\Delta_k}+\ldots+e^{\sum_{k=1}^{i}L(\tau_k)\Delta_k}\big)\cdot h(\tau_1). \tag{1.18}$$

Further, given that $h(\tau_1)=h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_1),\widetilde{X}_\beta^{\Gamma_*}(\tau_1))$ satisfies (1.7), from (1.18) we obtain the following estimate:

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}),\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant \big(1+e^{\sum_{k=i}^{i}L(\tau_k)\Delta_k}+e^{\sum_{k=i-1}^{i}L(\tau_k)\Delta_k}+ \\ +e^{\sum_{k=i-2}^{i}L(\tau_k)\Delta_k}+\ldots+e^{\sum_{k=1}^{i}L(\tau_k)\Delta_k}\big)\Delta\omega^*(\|\alpha-\beta\|). \tag{1.19}$$

Let us supplement the estimate (1.19) with a comment related to the function $L(t)$ continuous on the interval $[t_0,\vartheta]$, which was introduced in the condition **B**.

Remark 1. In numerous studies devoted to nonlinear control systems described by ordinary differential equations, the condition of the local Lipschitz property of its right-hand side with respect to the phase variable is introduced as one of the main conditions imposed on the system. In this case, often in the process of studying control problems for such systems, it becomes necessary to choose in the space of positions of the control system a domain D that would contain all the components of the resolving structure (resolving sets, trajectories of systems, phase constraints, etc.). In other words, quite often, when studying and solving control problems, it is necessary to choose a domain D in the space of positions of the system, in which the problem is solved. In this case, the Lipschitz constant L corresponding to this domain D is used for constructing a solution and justifying its correctness. However, the introduced domain D may turn out to be large, and the corresponding constant L may also turn out to be large. In this case, the estimates justifying the correctness of the solution of the control problem in which this constant L is involved may turn out to be rough. For various reasons, these estimates in a specific control problem (with a specific control system) may be unsatisfactory from the point of view of the person solving the problem and counting on finer estimates. In this regard, taking into consideration the conditions imposed on the nonlinear control system (1.1), in this paper, instead of the traditional local Lipschitz condition with the Lipschitz constant L, we introduce a continuous function $L(t)\in(0,\infty)$ on $[t_0,\vartheta]$, which is more suitable for the dynamics of (1.1). Estimate (1.19) of $h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}),\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1}))$ is more accurate in the sense that, for each interval $[\tau_i,\tau_{i+1}]$ of the partition Γ_*, the step-by-step estimates involve its own value $L(\tau_i) \in (0, \infty)$, which is close to $L(t)$, $t \in [\tau_i, \tau_{i+1}]$, for small $\Delta = \Delta(\Gamma_*)$, and not some constant $L \in (0, \infty)$ common to all $[\tau_i, \tau_{i+1}]$ from the interval $[t_0, \vartheta]$. Note, however, that this reasoning assumes that the domain D is in the position space of the system and the corresponding function $L(t)$ on $[t_0, \vartheta]$ is chosen sufficiently adequately to the dynamics of the control system. So, for example, in control problems related to the study of reachable sets and integral funnels, the domain D should track more or less accurately the dynamics of reachable sets and, therefore, the spatial structure of integral funnels.

Thus, in many specific control problems, the problem of choosing the domain D and the corresponding function $L(t)$, $t \in [t_0, \vartheta]$, in our opinion, is very significant, since the accuracy of the estimates related to solving problems depends on this.

Obviously, one of the ways to solve this problem in each specific task related to the study of reachable sets and integral funnels is to form the domain D and the function $L(t)$, $t \in [t_0, \vartheta]$, in a step-by-step procedure (by time layers $[\tau_i, \tau_{i+1}] \times \mathbb{R}^n$, $i = 0, 1, ..., N-1$) along with the construction of reachable sets.

Let us now return to estimate (1.19) and present some roughness of this estimate in a simpler form.

Replacing in (1.19) 1 and the exponents $e^{\sum_{k=r}^{i} L(\tau_k)\Delta_k}$, $r = \overline{1, i}$, by the exponent $e^{\sum_{k=0}^{i} L(\tau_k)\Delta_k}$, we get the estimate

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant e^{\sum_{k=0}^{i} L(\tau_k)\Delta_k} \cdot (i+1)\Delta\omega^*(\|\alpha - \beta\|),$$

i.e.,

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant e^{\sum_{k=0}^{i} L(\tau_k)\Delta_k} \cdot (\tau_{i+1} - \tau_0)\omega^*(\|\alpha - \beta\|). \tag{1.20}$$

In particular, the following estimate holds:

$$h(\widetilde{X}_\alpha^{\Gamma_*}(t^*), \widetilde{X}_\beta^{\Gamma_*}(t^*)) \leqslant e^{\sum_{k=0}^{N-1} L(\tau_k)\Delta_k} \cdot (t^* - t_*)\omega^*(\|\alpha - \beta\|). \tag{1.21}$$

Replacing in estimates (1.19)–(1.21) the numbers $L(\tau_k)$, $k = \overline{0, N-1}$, with some L satisfying the inequality $0 < \max\limits_{t \in [t_0, \vartheta]} L(t) \leqslant L < \infty$, we obtain the following estimates for $i \in \overline{1, N-1}$ and α, β from $\mathscr{L}$, respectively:

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant \sum_{k=0}^{i} e^{Lk\Delta}\Delta\omega^*(\|\alpha - \beta\|), \tag{1.22}$$

$$h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant e^{L \cdot (\tau_{i+1} - \tau_0)}(\tau_{i+1} - \tau_0)\omega^*(\|\alpha - \beta\|), \tag{1.23}$$

$$h(\widetilde{X}_\alpha^{\Gamma_*}(t^*), \widetilde{X}_\beta^{\Gamma_*}(t^*)) \leqslant e^{L \cdot (t^* - t_*)}(t^* - t_*)\omega^*(\|\alpha - \beta\|). \tag{1.24}$$

Reasoning similar to those given above for $h(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1})$ and $\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1}))$ yields estimates for $h(\widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}))$ similar to (1.19)–(1.24). Taking this into account, we come to the following statement.

Lemma 1. *Assume that* $[t_*, t^*] \subset [t_0, \vartheta]$, $X_* \in \operatorname{comp}(\mathbb{R}^n)$, $\Gamma_* = \{\tau_0 = t_*, \tau_1, ..., \tau_i, ..., \tau_N = t^*\}$ $(\tau_{i+1} - \tau_i = \Delta_i = \Delta$, $i = \overline{0, N-1})$, *and* $\{\widetilde{X}_\alpha^{\Gamma_*}(\tau_i) : \tau_i \in \Gamma_*\}$ *is the system of sets* (1.6) *approximating the reachable set* $X_\alpha(t^*, t_*, X_*)$, $\alpha \in \mathscr{L}$, *of the differential inclusion* (1.4). *Then, under the conditions* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *on system* (1.1), *the following estimates hold*:

$$d(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant \Big(1 + \sum_{s=0}^{i-1} e^{\sum_{k=i-s}^{i} L(\tau_k)\Delta_k}\Big)\Delta\omega^*(\|\alpha - \beta\|), \tag{1.25}$$

$$d(\widetilde{X}_\alpha^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_\beta^{\Gamma_*}(\tau_{i+1})) \leqslant e^{\sum_{k=0}^{i} L(\tau_k)\Delta_k}(\tau_{i+1} - \tau_0)\omega^*(\|\alpha - \beta\|), \tag{1.26}$$

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(t^*), \widetilde{X}_{\beta}^{\Gamma_*}(t^*)) \leqslant e^{\sum_{k=0}^{N-1} L(\tau_k)\Delta_k}(t^* - t_*)\omega^*(\|\alpha - \beta\|), \tag{1.27}$$

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_{\beta}^{\Gamma_*}(\tau_{i+1})) \leqslant \sum_{k=0}^{i} e^{Lk\Delta}\Delta\omega^*(\|\alpha - \beta\|), \tag{1.28}$$

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_{\beta}^{\Gamma_*}(\tau_{i+1})) \leqslant e^{L\cdot(\tau_{i+1}-\tau_0)} \cdot (\tau_{i+1} - \tau_0)\omega^*(\|\alpha - \beta\|), \tag{1.29}$$

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(t^*), \widetilde{X}_{\beta}^{\Gamma_*}(t^*)) \leqslant e^{L(t^*-t_*)}(t^* - t_*)\omega^*(\|\alpha - \beta\|). \tag{1.30}$$

From estimate (1.28), we derive estimates (1.29) and (1.30).

Let us write one more important estimate that follows from (1.28):

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(\tau_{i+1}, \widetilde{X}_{\beta}^{\Gamma_*}(\tau_{i+1})) \leqslant \frac{e^{(i+1)L\Delta} - 1}{e^{L\Delta} - 1}\Delta\omega^*(\|\alpha - \beta\|), \quad i = \overline{0, N-1}.$$

Let us estimate from above the right-hand side of this inequality, assuming that, along with the conditions **A** and **B** on system (1.1), the following condition on the partition Γ_* of the time interval $[t_*, t^*]$ holds.

C. The diameter of the partition Γ_* satisfies the relation

$$0 < \Delta = \Delta(\Gamma_*) < L^{-1} \ln\Big(1 + \frac{3}{2}L\Delta\Big).$$

Under the condition **C**, the following inequalities are valid:

$$\frac{e^{(i+1)L\Delta} - 1}{e^{L\Delta} - 1} < \frac{e^{L\Delta} \cdot e^{L(\tau_{i+1}-\tau_0)} - 1}{L\Delta}$$
$$< \frac{\Big(1 + 3/2 \cdot L\Delta\Big)e^{L(\tau_{i+1}-\tau_0)} - 1}{L\Delta} = \frac{e^{L(\tau_{i+1}-\tau_0)} - 1}{L\Delta} + \frac{3}{2}e^{L(\tau_{i+1}-\tau_0)}.$$

Taking this inequality into account, we obtain

$$\begin{aligned} d(\widetilde{X}_{\alpha}^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_{\beta}^{\Gamma_*}(\tau_{i+1})) < L^{-1}(e^{L(\tau_{i+1}-\tau_0)} - 1)\omega^*(\|\alpha - \beta\|)+ \\ +\frac{3}{2}e^{L(\tau_{i+1}-\tau_0)}\Delta\omega^*(\|\alpha - \beta\|), \quad i = \overline{0, N-1}. \end{aligned} \tag{1.31}$$

As a result, the following statement is true.

Theorem 1. *Let $[t_*, t^*] \subset [t_0, \vartheta]$ and $X_* \in \mathrm{comp}\,(\mathbb{R}^n)$. Then, under the conditions* ***A****,* ***B****, and* ***C*** *on system* (1.1)*, the sets $X_\alpha(t^*) = X_\alpha(t^*, t_*, X_*)$ and $X_\beta(t^*) = X_\beta(t^*, t_*, X_*)$, α and β from $\mathscr{L}$, satisfy estimate* (1.31)*.*

Obviously, for small $\Delta = \Delta(\Gamma_*)$, the strict estimate (1.31) will turn into the following estimate:

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(\tau_{i+1}), \widetilde{X}_{\beta}^{\Gamma_*}(\tau_{i+1})) \leqslant L^{-1}(e^{L(\tau_{i+1}-\tau_0)} - 1)\omega^*(\|\alpha - \beta\|), \quad i = \overline{0, N-1}.$$

In particular, the following statement is true.

Assertion 1. *Assume that $[t_*, t^*] \subset [t_0, \vartheta]$ and $X_* \in \mathrm{comp}\,(\mathbb{R}^n)$. Then, under the conditions* ***A****,* ***B****, and* ***C*** *on system* (1.1)*, the sets $X_\alpha(t^*) = X_\alpha(t^*, t_*, X_*)$ and $X_\beta(t^*) = X_\beta(t^*, t_*, X_*)$, α and β from $\mathscr{L}$, satisfy the estimate*

$$d(\widetilde{X}_{\alpha}^{\Gamma_*}(t^*), \widetilde{X}_{\beta}^{\Gamma_*}(t^*)) \leqslant L^{-1}(e^{L(t^*-t_*)} - 1)\omega^*(\|\alpha - \beta\|). \tag{1.32}$$

The question arises, at what ratios between the numbers L and $(t^* - t_*)$ one or another of estimates (1.30) and (1.32) is better. To answer it, let us compare the numbers $L^{-1}(e^{L(t^*-t_*)} - 1)$ and $e^{L(t^*-t_*)}(t^* - t_*)$, i.e., compare $e^{L(t^*-t_*)} - 1$ and $e^{L(t^*-t_*)}L(t^* - t_*)$.

Assuming that $L(t^* - t_*) = \rho > 0$, we come to the comparison of $e^\rho \cdot (1 - \rho)$ and 1 for $\rho \geqslant 0$.

Since the function $e^\rho \cdot (1 - \rho)$ equals 1 for $\rho = 0$ and decreases on $[0, \infty)$, we get

$$e^\rho \cdot (1 - \rho) < 1, \quad \rho > 0$$

and therefore

$$L^{-1}(e^{L(t^*-t_*)} - 1) < e^{L(t^*-t_*)}(t^* - t_*)$$

for every $L \in (0, \infty)$ and $(t^* - t_*) > 0$.

This means that estimate (1.32) is more precise than estimate (1.30) for sufficiently small $\Delta = \Delta(\Gamma_*)$.

We have considered the case $X_* = \{x_*\}$, $(t_*, x_*) \in D$, and received estimates (1.25)–(1.30). Estimates (1.25)–(1.30) are also true in the general case $X_* \in \mathrm{comp}\,(\mathbb{R}^n)$, $(t_*, X_*) \subset D$.

Bearing in mind the general case, we choose from (1.25)–(1.30) estimate (1.27) for the following reasons. Along with the sets $\widetilde{X}_\alpha^{\Gamma_*}(t^*)$ and $\widetilde{X}_\beta^{\Gamma_*}(t^*)$ included in (1.27), consider the reachable sets $X_\alpha(t^*) = X_\alpha(t^*, t_*, X_*)$ and $X_\beta(t^*) = X_\beta(t^*, t_*, X_*)$ of the differential inclusion (1.4).

We are looking for upper bounds for the values $d(X_\alpha(t^*), \widetilde{X}_\alpha^{\Gamma_*}(t^*))$ and $d(X_\beta(t^*),\ \widetilde{X}_\beta^{\Gamma_*}(t^*))$, where α and β are from $\mathscr{L}$. It is known that, under the conditions **A** and **B** on system (1.1), these estimates are of the form

$$\begin{aligned} d(X_\alpha(t^*), \widetilde{X}_\alpha^{\Gamma_*}(t^*)) &\leqslant e^{L\cdot(t^*-t_*)}(t^* - t_*)(\omega^*(\Delta) + LK\Delta), \\ d(X_\beta(t^*), \widetilde{X}_\beta^{\Gamma_*}(t^*)) &\leqslant e^{L\cdot(t^*-t_*)}(t^* - t_*)(\omega^*(\Delta) + LK\Delta); \end{aligned} \tag{1.33}$$

here $L \in (0, \infty)$ is defined on p. 127, $K = \max\limits_{(t,x,u,\alpha)\in D\times P\times\mathscr{L}} \|f_\alpha(t, x, u)\| \in (0, \infty)$, and $\Delta = \Delta(\Gamma_*)$.

Remark 2. It can be shown that, along with estimates (1.33), there are more subtle estimates:

$$\begin{aligned} d(X_\alpha(t^*), \widetilde{X}_\alpha^{\Gamma_*}(t^*)) &\leqslant e^{\int_{t_*}^{t^*} L(t)dt}(t^* - t_*)(\omega^*(\Delta) + LK\Delta), \\ d(X_\beta(t^*), \widetilde{X}_\beta^{\Gamma_*}(t^*)) &\leqslant e^{\int_{t_*}^{t^*} L(t)dt}(t^* - t_*)(\omega^*(\Delta) + LK\Delta). \end{aligned}$$

Taking into account (1.27) and (1.33), we get

$$\begin{aligned} d(X_\alpha(t^*), X_\beta(t^*)) &\leqslant d(X_\alpha(t^*), \widetilde{X}_\alpha^{\Gamma_*}(t^*)) + d(\widetilde{X}_\alpha^{\Gamma_*}(t^*), \widetilde{X}_\beta^{\Gamma_*}(t^*)) + d(\widetilde{X}_\beta^{\Gamma_*}(t^*), X_\beta(t^*)) \\ &\leqslant e^{\sum_{k=1}^{N-1} L(\tau_k)\Delta_k} \cdot (t^* - t_*)\omega^*(\|\alpha - \beta\|) + 2e^{L\cdot(t^*-t_*)} \cdot (\omega^*(\Delta) + LK\Delta), \end{aligned}$$

where α and β from $\mathscr{L}$.

Since this estimate holds for any partitions Γ_* of the interval $[t_*, t^*]$, letting the diameter $\Delta = \Delta(\Gamma_*)$ of the partition Γ_* tend to zero, we obtain

$$d(X_\alpha(t^*), X_\beta(t^*)) \leqslant e^{\int_{t_*}^{t^*} L(t)dt} \cdot (t^* - t_*) \cdot \omega^*(\|\alpha - \beta\|); \tag{1.34}$$

here $\int_{t_*}^{t^*} L(t)dt$ is the Riemann integral of the function $L(t)$ over the interval $[t_*, t^*] \subset [t_0, \vartheta]$.

Now let us turn to the interval $[t_0, \vartheta]$, on which the control system (1.1) and the differential inclusion (1.4) are initially considered.

Assume that in the previous calculations $t_* = t_0$, $t^* = t \in [t_0, \vartheta]$, $X_* = X^{(0)} \in \operatorname{comp}(\mathbb{R}^n)$, and $(t_0, X^{(0)}) \subset D$, where $X^{(0)}$ is the initial set for system (1.1) and the differential inclusion (1.4), so that the reachable sets $X_\alpha(t)$ and $X_\beta(t)$ of the differential inclusion (1.4) become $X_\alpha(t) = X_\alpha(t, t_0, X^{(0)})$ and $X_\beta(t) = X_\beta(t, t_0, X^{(0)})$.

For these sets, we write estimate (1.34):

$$d(X_\alpha(t), X_\beta(t)) \leqslant e^{\int_{t_0}^{t} L(\tau)d\tau} \cdot (t - t_0)\omega^*(\|\alpha - \beta\|), \tag{1.35}$$

where $t \in [t_0, \vartheta]$ and $\alpha, \beta \in \mathscr{L}$.

We also introduce the partition $\Gamma = \{t_0, t_1, ..., t_i, ..., t_N = \vartheta\}$ of the interval $[t_0, \vartheta]$ with the diameter $\Delta = \Delta(\Gamma) = t_{i+1} - t_i = N^{-1}(\vartheta - t_0)$.

Along with the reachable sets $X_\alpha(t)$, $\alpha \in \mathscr{L}$, $t \in [t_0, \vartheta]$, we consider the integral funnel

$$X_\alpha(t_0, X^{(0)}) = \bigcup_{t \in [t_0, \vartheta]} (t, X_\alpha(t)), \quad \alpha \in \mathscr{L},$$

of the differential inclusion (1.4).

Assume that

$$X_\alpha^\Gamma(t_0, X^{(0)}) = \bigcup_{t_i \in \Gamma} (t_i, X_\alpha(t_i)), \widetilde{X}_\alpha^\Gamma(t_0, X^{(0)}) = \bigcup_{t_i \in \Gamma} (t_i, \widetilde{X}_\alpha^\Gamma(t_i))$$

are sets in D, where $\widetilde{X}_\alpha^\Gamma(t_i)$ are defined on p. 122 by the recurrent relations with $\tau_0 = t_0$ and $\widetilde{X}_\alpha^\Gamma(t_0) = \widetilde{X}_\alpha^\Gamma(\tau_0) = X^{(0)}$.

Here the sets $X_\alpha^\Gamma(t_0, X^{(0)})$ and $\widetilde{X}_\alpha^\Gamma(t_0, X^{(0)})$ are some approximations of the integral funnel $X_\alpha(t_0, X^{(0)})$, $\alpha \in \mathscr{L}$, discrete by the parameter $t \in [t_0, \vartheta]$.

From the estimate

$$d(X_\alpha(t_i), \widetilde{X}_\alpha^\Gamma(t_i)) \leqslant e^{L\cdot(t_i - t_0)}(t_i - t_0)(\omega^*(\Delta) + LK\Delta), \quad i = \overline{1, N}, \quad \alpha \in \mathscr{L},$$

we obtain the estimate

$$d(X_\alpha^\Gamma(t_0, X^{(0)}), \widetilde{X}_\alpha^\Gamma(t_0, X^{(0)})) \leqslant e^{L\cdot(\vartheta - t_0)}(\vartheta - t_0)(\omega^*(\Delta) + LK\Delta); \tag{1.36}$$

here L is defined on p. 127 and K on p. 129.

Since the following inequality holds for each interval $[t_i, t_{i+1}]$ of the partition Γ, every $t \in [t_i, t_{i+1}]$, and every $\alpha \in \mathscr{L}$:

$$d\big((t, X_\alpha(t)), (t_i, X_\alpha(t_i))\big) \leqslant (1 + K)\Delta,$$

we have

$$d(X_\alpha(t_0, X^{(0)}), X_\alpha^\Gamma(t_0, X^{(0)})) \leqslant (1 + K)\Delta. \tag{1.37}$$

Considering estimates (1.36) and (1.37), we get

$$d(X_\alpha(t_0, X^{(0)}), \widetilde{X}_\alpha^\Gamma(t_0, X^{(0)})) \leqslant e^{L\cdot(\vartheta - t_0)}(\vartheta - t_0)(\omega^*(\Delta) + LK\Delta) + (1 + K)\Delta. \tag{1.38}$$

Obviously, using the technique of obtaining estimates described above, we can replace estimate (1.38) with a more accurate one:

$$d(X_\alpha(t_0, X^{(0)}), \widetilde{X}_\alpha^\Gamma(t_0, X^{(0)})) \leqslant e^{\sum_{i=0}^{N-1} L(t_i)\Delta_i} \cdot (\vartheta - t_0)(\omega^*(\Delta) + LK\Delta) + (1 + K)\Delta,$$
$$\Delta_i = \Delta = \Delta(\Gamma), \quad i \in \overline{0, N-1}, \quad \alpha \in \mathscr{L}.$$

Inequality (1.35) implies the following statement for the integral funnels $X_\alpha(t_0, X^{(0)})$ and $X_\beta(t_0, X^{(0)})$.

Theorem 2. *Let the control system* (1.1) *satisfy the conditions* ***A****,* ***B****, and* ***C****, and let* $X^{(0)} \in \operatorname{comp}(\mathbb{R}^n)$*. Then the integral funnels* $X_\alpha(t_0, X^{(0)})$ *and* $X_\beta(t_0, X^{(0)})$ *satisfy the inequality*

$$d(X_\alpha(t_0, X^{(0)}), X_\beta(t_0, X^{(0)})) \leqslant e^{\int_{t_0}^{\vartheta} L(t)dt} \cdot (\vartheta - t_0)\omega^*(\|\alpha - \beta\|), \quad \alpha, \beta \in \mathscr{L}.$$

2. Problems of targeting integral funnels to target sets in $\mathbb{R}^2$

In this section, we restrict ourselves to considering system (1.1) and the differential inclusion (1.4) in the space $\mathbb{R}^2$. Let us study problems of targeting integral funnels $X_\alpha(t_0, x_0)$, $\alpha \in \mathscr{L}$, $x_0 \in X^{(0)}$, and their approximations $\widetilde{X}_\alpha^\Gamma(t_0, x_0)$ to target sets in $\mathbb{R}^2$. We formulate some of these problems using the concept of the area of a set in $\mathbb{R}^2$. In this regard, we will study questions concerning the approximate calculation of the areas of reachable sets $X_\alpha(t, t_0, x_0)$, $x_0 \in X^{(0)} \in \operatorname{comp}(\mathbb{R}^2)$, and sets associated with $X_\alpha(t, t_0, x_0)$. In this case, we use the estimates of the Hausdorff distances obtained in Section 1.

Let us start the study of targeting problems by considering the individual integral funnels $X_\alpha(t_0, X^{(0)})$, $X^{(0)} \in \operatorname{comp}(\mathbb{R}^2)$. Of course, the funnels $X_\alpha(t_0, x_0)$, $\alpha \in \mathscr{L}$, $x_0 \in X^{(0)}$, also belong to the class of these funnels. Thus, the estimates of the Hausdorff distances obtained for integral funnels $X_\alpha(t_0, X^{(0)})$, $\alpha \in \mathscr{L}$, also hold for funnels $X_\alpha(t_0, x_0)$, $\alpha \in \mathscr{L}$.

Let us take an arbitrary funnel $X_\alpha(t_0, X^{(0)})$, $\alpha \in \mathscr{L}$, $X^{(0)} \in \operatorname{comp}(\mathbb{R}^2)$, and its approximating set $\widetilde{X}_\alpha^\Gamma(t_0, X^{(0)}) = \bigcup\limits_{t_i \in \Gamma} (t_i, \widetilde{X}_\alpha^\Gamma(t_i))$ in D corresponding to the partition $\Gamma = \{t_0, t_1, ..., t_i, ..., t_N = \vartheta\}$ $(t_{i+1} - t_i = \Delta_i = \Delta = \Delta(\Gamma),\ i = \overline{0, N-1})$.

The mismatch between the time sections $X_\alpha(t_i)$ and $\widetilde{X}_\alpha^\Gamma(t_i)$, $t_i \in \Gamma$, of the sets $X_\alpha(t_0, X^{(0)})$ and $\widetilde{X}_\alpha^\Gamma(t_0, X^{(0)})$ is restricted by the estimate

$$d(X_\alpha(t_i), \widetilde{X}_\alpha^\Gamma(t_i)) \leqslant e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot \Big(K\Delta \sum_{j=0}^{N-1} L(t_j)\Delta_j + (t_j - t_0)\omega^*(\Delta)\Big). \tag{2.1}$$

Along with the set $\widetilde{X}_\alpha^\Gamma(t_0, X^{(0)})$ and its sections $\widetilde{X}_\alpha^\Gamma(t_i)$, $t_i \in \Gamma$, we consider the set $\widetilde{X}_\beta^\Gamma(t_0, X^{(0)})$, $\beta \in \mathscr{L}$, and its sections $\widetilde{X}_\beta^\Gamma(t_i)$, $t_i \in \Gamma$. The following estimate is valid:

$$d(\widetilde{X}_\alpha^\Gamma(t_i), \widetilde{X}_\beta^\Gamma(t_i)) \leqslant e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot (t_i - t_0)\omega^*(\|\alpha - \beta\|). \tag{2.2}$$

Estimates (2.1) and (2.2) implpy

$$d(X_\alpha(t_i), \widetilde{X}_\beta^\Gamma(t_i)) \leqslant \varkappa(\Delta, \|\alpha - \beta\|), \tag{2.3}$$

where

$$\varkappa(\Delta, \rho) = e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j}\Big((\vartheta - t_0)\omega^*(\rho) + (\vartheta - t_0)\omega^*(\Delta) + K\Delta \sum_{j=0}^{N-1} L(t_j)\Delta_j\Big),$$

$$\alpha, \beta \in \mathscr{L}, \quad t_i \in \Gamma, \quad \rho \in (0, \infty).$$

We will use estimates (2.1)–(2.3) for studying problems of targeting integral funnels to target sets. These estimates will also be taken into account when estimating the mismatch of sets of the type of reachable sets in $\mathbb{R}^2$.

Let us formulate these targeting problems.

Assume that a finite set $\mathcal{T}$ of times η_1, η_2, ..., η_{N_*} from the interval $[t_0, \vartheta]$ is given and the partition $\Gamma = \{t_0, t_1, ..., t_i, ..., t_N = \vartheta\}$ considered previously contains this set $\mathcal{T}$.

Assume that compact sets $X^{(0)}$, $X^{(\vartheta)}$, and $\Phi^{(k)}$ in $\mathbb{R}^2$ are given, where each set $\Phi^{(k)}$ corresponds to its time $\eta_k \in \mathcal{T}$; moreover, the sets $X^{(0)}$, $X^{(\vartheta)}$, and $\Phi^{(k)}$, $\eta_k \in \mathcal{T}$, have rectifiable boundaries $\partial X^{(0)}$, $\partial X^{(\vartheta)}$, and $\partial\Phi^{(k)}$, $\eta_k \in \mathcal{T}$.

Here we assume that $\Phi^{(k)} = \Phi(\eta_k)$, $\eta_k \in \mathcal{T}$, where the set $\Phi(t) \in \operatorname{comp}(\mathbb{R}^2)$, $t \in [t_0, \vartheta]$, is interpreted by us as an obstacle to system (1.1).

Problem 1 on targeting integral funnels (strict setting). It is required to find a pair $(\alpha_*, x_*) \in \mathscr{L} \times X^{(0)}$ such that the following relations hold:

$$X^{(\vartheta)} \subset X_{\alpha_*}(\vartheta, t_0, x_*), \quad \Phi^{(k)} \cap X_{\alpha_*}(\eta_k, t_0, x_*) = \varnothing, \quad \eta_k \in \mathcal{T}.$$

Exact computation of the sets $X_\alpha(t_i, t_0, x_0)$, $\alpha \in \mathscr{L}$, $t_i \in \Gamma$, $x_0 \in X^{(0)}$ is not possible due to the complexity of the system dynamics (1.1). In particular, it is impossible to compute the sets $X_{\alpha_*}(\vartheta, t_0, x_*)$ and $X_{\alpha_*}(\eta_k, t_0, x_*)$, $\eta_k \in \mathcal{T}$. Also in the case when, for example, one of the sets $\mathscr{L}$ and $X^{(0)}$ is infinite, the complete enumeration of all pairs $(\alpha, x) \in \mathscr{L} \times X^{(0)}$ is impossible.

Therefore, it makes sense to go from the statement of Problem 1 to a statement in terms of the sets $\widetilde{X}^{\Gamma}_\alpha(t_i, t_0, x_0)$, $\alpha \in \mathscr{L}$, $t_i \in \Gamma$, $x_0 \in X^{(0)}$. Moreover, under the sets $\widetilde{X}^{\Gamma}_\alpha(t_i, t_0, x_0)$ we understand time sections of the sets $\widetilde{X}^{\Gamma}_\alpha(t_0, x_0)$, $\alpha \in \mathscr{L}$, $x_0 \in X^{(0)}$, corresponding to the times $t_i \in \Gamma$.

More precisely, we assume that there are given ε, ρ, and σ from $(0, \infty)$ and finite sets corresponding to the numbers ρ and σ in the sets $\mathscr{L}$ and $X^{(0)}$, a ρ-net $\mathscr{L}^{(\rho)} = \{\alpha^{(r)} : r = \overline{1, r_*}\}$ and a σ-net $X^{(\sigma)} = \{x^{(s)} : s = \overline{1, s_*}\}$.

Problem $1^{(\varepsilon)}$ on targeting integral funnels. It is required to find a pair $(\alpha^{(r)}, x^{(s)}) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$ such that the following relations hold:

$$X^{(\vartheta)} \subset \widetilde{X}^{\Gamma}_{\alpha^{(r)}}(\vartheta, t_0, x^{(s)})_\varepsilon, \quad \Phi^{(k)}_\varepsilon \cap \widetilde{X}^{\Gamma}_{\alpha^{(r)}}(\eta_k, t_0, x^{(s)}) = \varnothing, \quad \eta_k \in \mathcal{T}.$$

For Problems 1 or $1^{(\varepsilon)}$ formulated for a particular system (1.1), it may turn out that there is no solution. Taking into account such situations, we formulate the targeting problem in a less strict setting, using the concept of the area of a set in $\mathbb{R}^2$. At the same time, we assume that such a formulation does not contradict the meaning of the original real targeting problem.

First, we give a statement in terms of ideal reachable sets $X_\alpha(t_i, t_0, x_0)$, $\alpha \in \mathscr{L}$, $x_0 \in X^{(0)}$, $t_i \in \Gamma$.

Let us introduce the notation

$$J^{(1)}(\alpha, x) = \sum_{\eta_k \in \mathcal{T}} s(\Phi^{(k)} \backslash X_\alpha(\eta_k, t_0, x)),$$
$$J^{(2)}(\alpha, x) = s(X^{(\vartheta)} \cap X_\alpha(\vartheta, t_0, x)),$$
$$\alpha \in \mathscr{L}, \quad x \in X^{(0)};$$

here $s(Y)$ is the area of the set $Y \in \mathrm{comp}\,(\mathbb{R}^2)$, by which we mean the Lebesgue measure (see, e.g., [4, Ch. 2, Sect. 2.5]) of the compact set Y in $\mathbb{R}^2$.

Let us fix λ_1 and λ_2 from $[0, 1]$, $\lambda_1 + \lambda_2 = 1$.

Let us clarify once again that, under a strict setting of the problem of targeting integral funnels of the differential inclusion (1.4), we mean a setting in which an integral funnel $X_\alpha(t_0, x^{(0)})$, $\alpha \in \mathscr{L}$, should not intersect an obstacle $\Phi(t)$, $t \in [t_0, \vartheta]$; in the worst case, it can only touch its boundary $\partial\Phi(t)$, $t \in [t_0, \vartheta]$. In this case, the integral funnel $X_\alpha(t_0, x^{(0)})$, $\alpha \in \mathscr{L}$, must completely cover the target set $X^{(\vartheta)}$ at the terminal time ϑ.

The soft setting of the targeting problem allows the integral funnel $X_\alpha(t_0, x^{(0)})$ to creep on the obstacle $\Phi(t)$, $t \in [t_0, \vartheta]$ and admits incomplete coverage of the target set $X^{(\vartheta)}$ by the integral funnel $X_\alpha(t_0, x^{(0)})$ (more precisely, by its latter section $X_\alpha(\vartheta, t_0, x^{(0)})$) at the time ϑ. However, this involves some quantitative estimates of the effectiveness of the integral funnel $X_\alpha(t_0, x^{(0)})$ when solving the problem of targeting $X^{(\vartheta)}$. These quantitative estimates are associated with calculating the areas of sets in the space $\mathbb{R}^2$.

We assume that $J(\alpha, x) = \lambda_1 J^{(1)}(\alpha, x) + \lambda_2 J^{(2)}(\alpha, x)$.

Problem 2 on targeting integral funnels (soft setting). It is required to find a pair $(\alpha^*, x^*) \in \mathscr{L} \times X^{(0)}$ such that the following relation is true:

$$J(\alpha^*, x^*) = \max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(\alpha, x). \tag{2.4}$$

Since we are not able to solve Problem 2 exactly for the same reasons as Problem 1, we formulate and solve some approximation problem in which, instead of the sets $\mathscr{L}$ and $X^{(0)}$, in the cases where they are not finite, there are their finite nets $\mathscr{L}^{(\rho)}$ and $X^{(\sigma)}$ and, instead of (ideal) reachable sets $X_\alpha(t, t_0, x_0)$, $\alpha \in \mathscr{L}$, $x_0 \in X^{(0)}$, there are their approximations $\widetilde{X}^{\Gamma}_{\alpha^{(r)}}(t_i, t_0, x^{(s)})$, $(\alpha^{(r)}, x^{(s)}) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$.

Let us introduce the notation

$$\widetilde{J}^{(1)}_{\Gamma}(\beta, y) = \sum_{\eta_k\in\mathcal{T}} s(\Phi^{(k)} \backslash \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, y)),$$

$$\widetilde{J}^{(2)}_{\Gamma}(\beta, y) = s(X^{(\vartheta)} \cap \widetilde{X}^{\Gamma}_{\beta}(\vartheta, t_0, y)),$$

$$(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}.$$

We assume that $\widetilde{J}_{\Gamma}(\beta, y) = \lambda_1 \widetilde{J}^{(1)}_{\Gamma}(\beta, y) + \lambda_2 \widetilde{J}^{(2)}_{\Gamma}(\beta, y)$.

Problem 3 on targeting integral funnels (soft setting). It is required to find a pair $(\beta^*, y^*) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$ such that the following relation holds:

$$\widetilde{J}_{\Gamma}(\beta^*, y^*) = \max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}} \widetilde{J}_{\Gamma}(\beta, y). \tag{2.5}$$

Let us show that, for small ρ and σ from $(0, \infty)$, the solution of the approximation Problem 3 is close to the solution of Problem 2. This circumstance justifies replacing Problem 2 with Problem 3. In this case, we understand the proximity of solutions as the proximity of optimal values (2.4) and (2.5) in Problems 2 and 3 and the proximity of optimal pairs in $\mathscr{L} \times X^{(0)}$ and $\mathscr{L}^{(\rho)} \times X^{(\sigma)}$.

So, consider first pairs (α, x) and (β, y), where (α, x) is chosen in $\mathscr{L} \times X^{(0)}$ arbitrarily and the pair $(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$ is such that $\|\alpha - \beta\| \leqslant \rho$ and $\|x - y\| \leqslant \sigma$.

Let us find an upper bound for the Hausdorff distance

$$d(X_\alpha(\eta_k, t_0, x), \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, y)), \quad \eta_k \in \mathcal{T}.$$

In view of (2.3) and the estimate

$$d(\widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, x), \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, y)) \leqslant e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot \|x - y\| \leqslant e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot \sigma,$$

we have

$$\begin{aligned} d(X_\alpha(\eta_k, t_0, x), \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, y)) &\leqslant d(X_\alpha(\eta_k, t_0, x), \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, x)) + d(\widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, x), \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, y)) \\ &\leqslant \varkappa(\Delta, \rho) + e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot \sigma, \quad \eta_k \in \mathcal{T}. \end{aligned}$$

For simplicity, we introduce the notation

$$\varkappa^{\Delta}(\rho, \sigma) = \varkappa(\Delta, \rho) + e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot \sigma, \quad \rho, \sigma \in (0, \infty).$$

Finally, for pairs $(\alpha, x) \in \mathscr{L} \times X^{(0)}$ and $(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$ such that $\|\alpha - \beta\| \leqslant \rho$ and $\|x - y\| \leqslant \sigma$, we have the estimate

$$d(X_\alpha(\eta_k, t_0, x), \widetilde{X}^{\Gamma}_{\beta}(\eta_k, t_0, y)) \leqslant \varkappa^{\Delta}(\rho, \sigma). \tag{2.6}$$

Let us describe the function $\varkappa^{\Delta}(\rho,\sigma)$ in more detail and estimate it from above. The following representation is valid:

$$\varkappa^{\Delta}(\rho,\sigma) = e^{\sum_{j=0}^{N-1} L(t_j)\Delta_j} \cdot \Big((\vartheta - t_0)\omega^*(\rho) + (\vartheta - t_0)\omega^*(\Delta) + K\Delta \sum_{j=0}^{N-1} L(t_j)\Delta_j + \sigma\Big).$$

Since, by the condition **A**, the function $L(t) \in (0,\infty)$ is continuous on $[t_0,\vartheta]$, the following estimate is valid for $L \in \big(\max\limits_{t\in[t_0,\vartheta]} L(t), \infty\big)$:

$$\varkappa^{\Delta}(\rho,\sigma) \leqslant e^{L\cdot(\vartheta-t_0)}\big((\vartheta - t_0)\omega^*(\rho) + (\vartheta - t_0)\omega^*(\Delta) + LK(\vartheta - t_0)\Delta + \sigma\big).$$

This estimate implies the limit equality $\lim\limits_{\Delta\downarrow 0,\rho\downarrow 0,\sigma\downarrow 0} \varkappa^{\Delta}(\rho,\sigma) = 0$.

We supplement the conditions **A**, **B** and **C** with the following condition.

D. The lengths of the boundaries $\partial X^{(0)}$, $\partial X^{(\vartheta)}$, $\partial\Phi^{(k)}$, $\partial X_\alpha(t_i,t_0,x)$, and $\partial \widetilde{X}^{\Gamma}_{\beta}(t_i,t_0,y)$ ($(\alpha,x) \in \mathscr{L} \times X^{(0)}$, $(\beta,y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$, $\eta_k \in \mathcal{T}$, $t_i \in \Gamma$) are bounded from above by some $l^* \in (0,\infty)$.

The condition **D** holds for many problems on guiding integral funnels, since the lengths of the boundaries $\partial X^{(0)}$, $\partial X^{(\vartheta)}$, and $\partial \Phi^{(k)}$ ($\eta_k \in \mathcal{T}$) are bounded, and the lengths of the boundaries $\partial X_\alpha(t_i,t_0,x)$ and $\partial \widetilde{X}^{\Gamma}_{\beta}(t_i,t_0,y)$, $t_i \in \Gamma$, do not increase abruptly with increasing the times t_i. So, for example, the set $X_\alpha(t,t_0,x)$, $\alpha \in \mathscr{L}$, $x \in X^{(0)}$, continuously depends on t on $[t_0,\vartheta]$ (see Sect. 1, p. 122) and the set $\partial X_\alpha(t,t_0,x)$ also continuously depends on t on $[t_0,\vartheta]$ for many control problems. In these problems, it continuously depends on t and the length of the boundary $\partial X_\alpha(t,t_0,x)$.

Let

$$U_\alpha(\eta_k) = \mathrm{cl}(X_\alpha(\eta_k,t_0,x)_{\varkappa^{\Delta}(\rho,\sigma)} \backslash X_\alpha(\eta_k,t_0,x))$$

be the $\varkappa^{\Delta}(\rho,\sigma)$-layer around the set $X_\alpha(\eta_k,t_0,x)$, and let

$$\widetilde{U}^{\Gamma}_{\beta}(\eta_k) = \mathrm{cl}\big((\widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y)_{\varkappa^{\Delta}(\rho,\sigma)} \backslash \widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y)\big)$$

be the $\varkappa^{\Delta}(\rho,\sigma)$-layer around the set $\widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y)$.

Estimate (2.6) implies

$$\begin{aligned} X_\alpha(\eta_k,t_0,x) &\subset \widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) \cup \widetilde{U}^{\Gamma}_{\beta}(\eta_k), \\ \widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) &\subset X_\alpha(\eta_k,t_0,x) \cup U_\alpha(\eta_k). \end{aligned} \tag{2.7}$$

From inclusions (2.7), we obtain

$$\begin{aligned} X_\alpha(\eta_k,t_0,x) \cap \Phi^{(k)} &\subset \big(\widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) \cap \Phi^{(k)}\big) \cup \big(\widetilde{U}^{\Gamma}_{\beta}(\eta_k) \cap \Phi^{(k)}\big), \quad \eta_k \in \mathcal{T}, \\ \widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) \cap \Phi^{(k)} &\subset \big(X_\alpha(\eta_k,t_0,x) \cap \Phi^{(k)}\big) \cup \big(U_\alpha(\eta_k) \cap \Phi^{(k)}\big), \quad \eta_k \in \mathcal{T}. \end{aligned} \tag{2.8}$$

Inclusions (2.8) imply the following inequalities for the areas:

$$\begin{aligned} s(X_\alpha(\eta_k,t_0,x) \cap \Phi^{(k)}) &\leqslant s(\widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) \cap \Phi^{(k)}) + s(\widetilde{U}^{\Gamma}_{\beta}(\eta_k)), \quad \eta_k \in \mathcal{T}, \\ s(\widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) \cap \Phi^{(k)}) &\leqslant s(X_\alpha(\eta_k,t_0,x) \cap \Phi^{(k)}) + s(U_\alpha(\eta_k)), \quad \eta_k \in \mathcal{T}. \end{aligned} \tag{2.9}$$

From inequalities (2.9), we derive the estimate

$$\begin{gathered} \big|s(X_\alpha(\eta_k,t_0,x) \cap \Phi^{(k)}) - s(\widetilde{X}^{\Gamma}_{\beta}(\eta_k,t_0,y) \cap \Phi^{(k)})\big| \leqslant \\ \leqslant \max\big(s(U_\alpha(\eta_k), s(\widetilde{U}^{\Gamma}_{\beta}(\eta_k))\big), \quad \eta_k \in \mathcal{T}. \end{gathered} \tag{2.10}$$

Let us make a short note about the layers surrounding compact sets in $\mathbb{R}^2$; these layers include the sets $U_\alpha(\eta_k)$ and $\widetilde{U}^\Gamma_\beta(\eta_k)$, $\eta_k \in \mathcal{T}$.

It is known (see., e.g., [15]) that if $X \in \mathrm{comp}\,(\mathbb{R}^2)$ is a convex set, then the area $s(U_\varepsilon)$ of the ε-layer $U_\varepsilon = \mathrm{cl}(X_\varepsilon \backslash X)$ surrounding X and the length $l(\partial X)$ of the boundary ∂X of the set X are connected as follows:

$$s(U_\varepsilon) = l(\partial X) \cdot \varepsilon + \pi \cdot \varepsilon^2. \tag{2.11}$$

If the set $X \in \mathrm{comp}\,(\mathbb{R}^2)$ is not convex and connected, then the area $s(U_\varepsilon)$ can satisfy the inequality

$$s(U_\varepsilon) \leqslant l(\partial X) \cdot \varepsilon + \pi \cdot \varepsilon^2, \tag{2.12}$$

which we will use to estimate the areas $s(U_\alpha(\eta_k), s(\widetilde{U}^\Gamma_\beta(\eta_k))$, $\tau_k \in \mathcal{T}$.

Remark 3. We will give examples of non-convex sets for which equality (2.11) turns into inequality (2.12) and examples of non-convex sets for which equality (2.11) is satisfied. We will also demonstrate that the connectedness condition is necessary.

Example 1. Consider the simplest example of a convex set: $X = \{x = (x_1, x_2) : \|x\| \leqslant R\}$, where $R > 0$ is the radius of the disk X (Fig. 1).

In this case, equality (2.11) is easily verified by direct computation. Indeed,

$$s(U_\varepsilon) = \pi(R+\varepsilon)^2 - \pi R^2 = 2\pi R\varepsilon + \pi\varepsilon^2.$$

Here the length of the boundary ∂X is equal to $l(\partial X) = 2\pi R$ in full accordance with (2.11).

Example 2. Consider a non-convex set (Fig. 2)

$$X = \big\{x = (x_1, x_2) : \max\{|x_1|, |x_2|\} \leqslant R,$$
$$\|x - A^*\| > R,\ \|x - B^*\| > R,\ \|x - C^*\| > R,\ \|x - D^*\| > R\big\},$$

where $A^* = (-R, -R)$, $B^* = (-R, R)$, $C^* = (R, R)$, and $D^* = (R, -R)$, $R > 0$.

We denote by $K = \big\{x = (x_1, x_2) : \max\{|x_1|, |x_2|\} \leqslant R\big\}$ the square $A^*B^*C^*D^*$. In this case, the ε-layer U_ε consists of four semidisks of radius ε centered in the middle of the segments A^*B^*, B^*C^*, C^*D^*, and A^*D^* and the four curvilinear sets

$$U_1 = (B(A^*, R) \backslash B(A^*, R-\varepsilon)) \cap K,$$
$$U_2 = (B(B^*, R) \backslash B(B^*, R-\varepsilon)) \cap K,$$
$$U_3 = (B(C^*, R) \backslash B(C^*, R-\varepsilon)) \cap K,$$

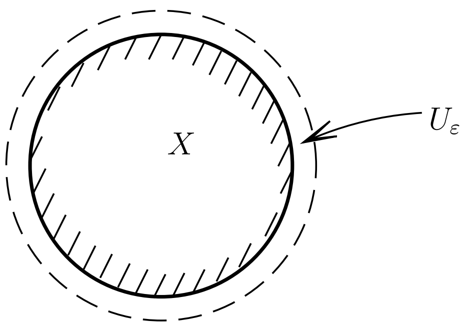


Figure 1. Example 1: the simplest convex set in the form of a disk.

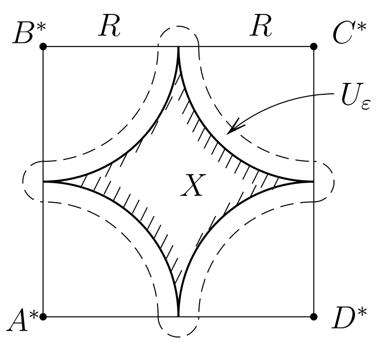


Figure 2. Example 2: a simple non-convex set.

$$U_4 = (B(D^*, R) \backslash B(D^*, R - \varepsilon)) \cap K,$$

where $B(a, r) = \{x = (x_1, x_2) : \|x - a\| \leqslant r\}$ denotes the closed disk of radius $r > 0$ centered at a point $a \in \mathbb{R}^2$.

Obviously, the total area of the four semidisks is $2\pi\varepsilon^2$, and the sum of the areas of the curvilinear sets is

$$s(U_1) + s(U_2) + s(U_3) + s(U_4) = \pi R^2 - \pi(R - \varepsilon)^2 = 2\pi R\varepsilon - \pi\varepsilon^2.$$

As a result, we get

$$s(U_\varepsilon) = 2\pi R\varepsilon + \pi\varepsilon^2.$$

Since $l(\partial X) = 2\pi R$, equality (2.11) is satisfied in this case even though X is a non-convex set.

Example 3. Consider another non-convex set $X = X_1 \cup X_1$, (Fig. 3), where

$$X_1 = \{x = (x_1, x_2) : R \leqslant \|x - O_1\| \leqslant R + \mu,\ x_2 \geqslant 0\},$$
$$X_2 = \{x = (x_1, x_2) : R \leqslant \|x - O_2\| \leqslant R + \mu,\ x_2 \leqslant 0\},$$
$$O_1 = (0, 0), \quad O_2 = (2R + \mu, 0), \quad R > 0, \quad \mu > 0.$$

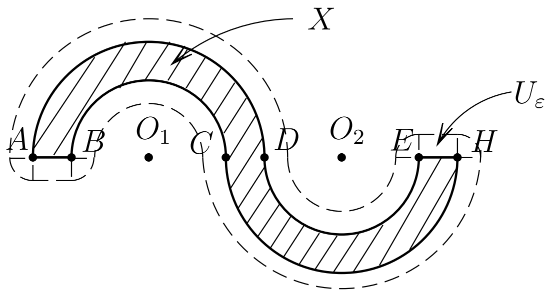


Figure 3. Example 3: a non-convex S-shaped set.

In this case, the ε-layer consists of four quarters of disk of radius ε centered at the points A, B, E, and H, respectively, and four curvilinear sets adjacent to the arcs AD, BC, DE, and CH.

It is easy to calculate that, firstly,

$$l(\partial X) = 2\mu + 2\pi R + 2\pi(R + \mu);$$

secondly,

$$s(U_\varepsilon) = \pi\varepsilon^2 + \pi(R + \mu + \varepsilon)^2 - \pi(R + \mu)^2 + \pi R^2 - \pi(R - \varepsilon)^2 = \pi\varepsilon^2 + 2\pi(R + \mu)\varepsilon + 2\pi R\varepsilon.$$

Thus, in this case, equality (2.11) holds despite the non-convexity of X.

Example 4. Let us give an example of a non-convex set X for which equality (2.11) nevertheless turns into inequality (2.12). Let $X = K_2 \backslash K_1$ (Fig. 4), where

$$K_1 = \{x = (x_1, x_2) : \max\{|x_1|, |x_2|\} \leqslant 1\},$$
$$K_2 = \{x = (x_1, x_2) : \max\{|x_1|, |x_2|\} \leqslant 2\}.$$

In this case, it is easy to calculate that

$$l(\partial X) = 8 + 4 = 12, \quad s(U_\varepsilon) = 8\varepsilon + \pi\varepsilon^2 + 4\varepsilon - 4\varepsilon^2 = 12\varepsilon - (4 - \pi)\varepsilon^2,$$

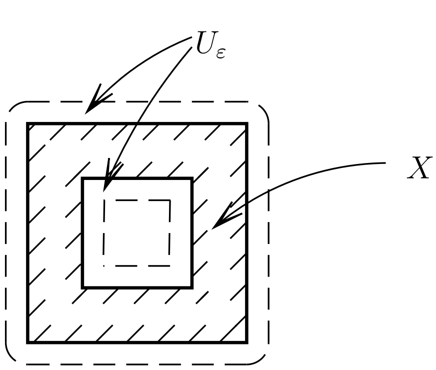


Figure 4. Example 4: a set with a cut-out square hole for which equality (2.11) is violated.

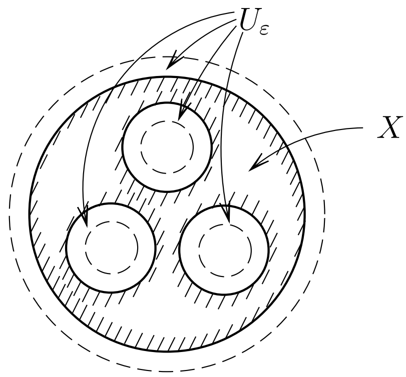


Figure 5. Example 5: a set with three holes for which equality (2.11) is violated to a large extent.

i.e., instead of equality (2.11), inequality (2.12) holds.

Example 5. Obviously, by increasing the number of "holes" inside the set X, one can increase the absolute value of the difference between $s(U_\varepsilon)$ and $l(\partial X)\cdot\varepsilon+\pi\varepsilon^2$. For example, the area of the ε-layer for the set X shown in Fig. 5 (and consisting of a disk with three round holes)

$$s(U_\varepsilon)=l(\partial X)\cdot\varepsilon-2\pi\varepsilon^2$$

regardless of the values of the radii of the disk and holes, provided that the radii of all holes are not less than ε.

Example 6. Note that the presence of "holes" inside X is not necessary to violate equality (2.11). Fig. 6 shows a simply connected set X for which

$$s(U_\varepsilon)=l(\partial X)\cdot\varepsilon+5\pi\varepsilon^2-16\varepsilon^2.$$

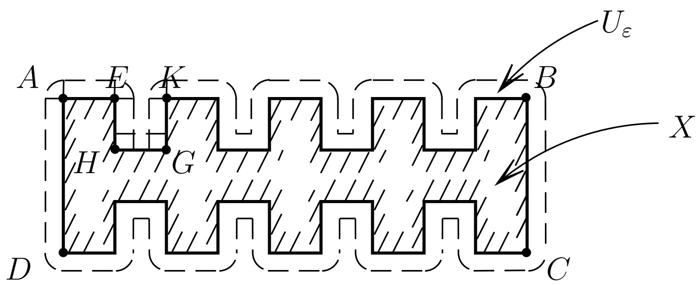


Figure 6. Example 6: a simply connected set with a negative quadratic addition in the formula for the ε-layer area.

Example 7. This example of the set X consisting of two separate disks of radius R (Fig. 7) shows that the connectedness condition for the set X is necessary for inequality (2.12). Indeed, in this case,

$$s(U_\varepsilon)=l(\partial X)\cdot\varepsilon+2\pi\varepsilon^2,$$

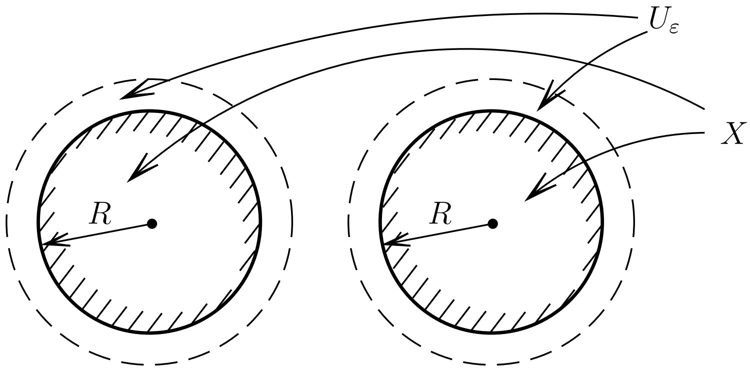


Figure 7. Example 7: a non-simply connected set for which inequality (2.12) is violated.

which violates inequality (2.12).

Example 8. In the last example, we will show that the addition to the main part $l(\partial X) \cdot \varepsilon$ in the expression for $s(U_\varepsilon)$ is not always proportional to ε^2. Indeed, let $X = B(O_1, R) \cup B(O_2, R)$, $0 < \varepsilon < R$. In this case, the boundaries $\partial B(O_1, R)$ and $\partial B(O_2, R)$ (which are circles) intersect at two points B and D. Define $\angle O_1O_2B = \varphi$ (Fig. 8).

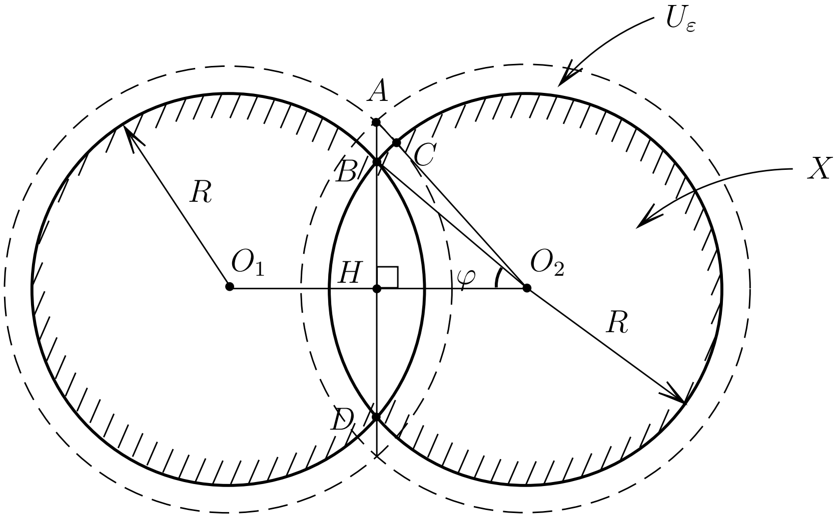


Figure 8. Example 8: two intersecting disks.

Then $l(\partial X) = 4R(\pi - \varphi)$.
From ΔBHO_2, we get $|BH| = R\sin\varphi, \quad |HO_2| = R\cos\varphi$.
Considering the right-angled triangle ΔAHO_2 and the Pythagorean theorem, we find that

$$|AO_2| = R + \varepsilon, \quad |AH| = \sqrt{(R+\varepsilon)^2 - R^2\cos^2\varphi}.$$

Further, the length of the segment AB is

$$|AB| = |AH| - |BH| = \sqrt{(R+\varepsilon) - R^2\cos^2\varphi} - R\sin\varphi,$$

the area of the triangle ΔABO_2 is

$$s(\Delta ABO_2) = \frac{1}{2}|AB| \cdot |HO_2| = \frac{1}{2}\big(\sqrt{(R+\varepsilon) - R^2\cos^2\varphi} - R\sin\varphi\big)R\cos\varphi,$$

the value of the angle $\angle HO_2A$ is

$$\angle HO_2A = \arccos\frac{|HO_2|}{|AO_2|} = \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big),$$

and the value of the angle $\angle BO_2A$ is

$$\angle BO_2A = \angle HO_2A - \varphi = \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big) - \varphi.$$

Consider the figure ABC whose sides AB and $AC \subset AO_2$ are segments and BC is an arc of the circle centered at O_2. Denote by O_2BC the sector based on the arc BC. The area of the figure ABC is

$$\begin{gathered} s(ABC) = s(\Delta ABO_2) - s(O_2BC) \\ = \frac{1}{2}\big(\sqrt{(R+\varepsilon)^2 - R^2\cos^2\varphi} - R\sin\varphi\big)R\cos\varphi - \frac{1}{2}R^2\Big(\arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big) - \varphi\Big). \end{gathered}$$

Taking into account the symmetry of the set X about the lines AD and O_1O_2, we obtain the area of the ε-layer:

$$\begin{gathered} s(U_\varepsilon) = 4s(ABC) + 2\cdot\frac{2\pi - 2\varphi}{2}\big((R+\varepsilon)^2 - R^2\big) \\ = 2\big(\sqrt{(R+\varepsilon)^2 - R^2\cos^2\varphi} - R\sin\varphi\big)R\cos\varphi \\ +2R^2\Big(\varphi - \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big)\Big) + 2\Big(\pi - \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big)\Big)(2R\varepsilon + \varepsilon^2) \end{gathered}$$

or

$$\begin{gathered} s(U_\varepsilon) = l(\partial X)\cdot\varepsilon + 4\Big(\varphi - \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big)\Big)R\varepsilon \\ +2\big(\sqrt{(R+\varepsilon)^2 - R^2\cos^2\varphi} - R\sin\varphi\big)R\cos\varphi \\ +2R^2\Big(\varphi - \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big)\Big) + 2\Big(\pi - \arccos\Big(\frac{R\cos\varphi}{R+\varepsilon}\Big)\Big)\varepsilon^2. \end{gathered}$$

Note that, in the particular case $\varphi = 0$,

$$\begin{gathered} s(U_\varepsilon) = l(\partial X)\cdot\varepsilon - 4R\arccos\Big(\frac{R}{R+\varepsilon}\Big)\varepsilon \\ +2\big(\sqrt{(R+\varepsilon)^2 - R^2}\big)R - 2R^2\arccos\Big(\frac{R}{R+\varepsilon}\Big) + 2\Big(\pi - \arccos\Big(\frac{R}{R+\varepsilon}\Big)\Big)\varepsilon^2 \\ = l(\partial X)\cdot\varepsilon - \frac{8}{3}\sqrt{2R}\cdot\varepsilon^{3/2} + 2\pi\varepsilon^2 - \frac{14}{15}\frac{\sqrt{2}\varepsilon^{5/2}}{\sqrt{R}} + \frac{71}{420}\frac{\sqrt{2}\varepsilon^{7/2}}{R^{3/2}} + O(\varepsilon^{9/2}), \quad \varepsilon \to 0. \end{gathered}$$

Thus, the addition to $l(\partial X)$ in the expression for $s(U_\varepsilon)$ may have rather complicated asymptotics and may not start with a term of the form $C\cdot\varepsilon^2$.

Remark 4. As can be seen from the examples, the question of the asymptotic behavior of the areas of the ε-layers U_ε surrounding the set X is nontrivial. It is related to questions of the geometric and topological structure of sets X in $\mathbb{R}^2$ and has an independent meaning.

In connection with Remark 3, we introduce one more condition concerning the sets $X_\alpha(t_i, t_0, x)$ and $\widetilde{X}^\Gamma_\beta(t_i, t_0, y)$, $(\alpha, x) \in \mathscr{L} \times X^{(0)}$, $(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$, $t_i \in \Gamma$.

E. The areas $s(U_\alpha(t_i))$, $s(\widetilde{U}^\Gamma_\beta(t_i))$, $\alpha, \beta \in \mathscr{L}$, $t_i \in \Gamma$, satisfy inequality (2.12) for $\varepsilon = \varkappa^\Delta(\rho, \sigma)$:

$$s(U_\alpha(t_i)) \leqslant l(\partial X_\alpha(t_i, t_0, x))\varepsilon + \pi\varepsilon^2,$$
$$s(\widetilde{U}_\beta(t_i)) \leqslant l(\partial \widetilde{X}^\Gamma_\beta(t_i, t_0, y))\varepsilon + \pi\varepsilon^2.$$

Taking into account the definition of the sets $U_\alpha(\eta_k)$ and $\widetilde{U}^\Gamma_\beta(\eta_k)$, $\eta_k \in \mathcal{T}$, and the condition **E**, we obtain

$$\begin{gathered} \max\big(s(U_\alpha(\eta_k)), s(\widetilde{U}^\Gamma_\beta(\eta_k))\big) \leqslant \max\big(l(\partial X_\alpha(\eta_k, t_0, x) \cap \Phi^{(k)}), \\ l(\partial \widetilde{X}^\Gamma_\beta(\eta_k, t_0, y) \cap \Phi^{(k)})\big) \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2 \leqslant l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2. \end{gathered} \tag{2.13}$$

From (2.10) and (2.13), it follows that

$$\big|s(X_\alpha(\eta_k, t_0, x) \cap \Phi^{(k)}) - s(\widetilde{X}^\Gamma_\beta(\eta_k, t_0, y) \cap \Phi^{(k)})\big| \leqslant l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2, \quad \eta_k \in \mathcal{T}. \tag{2.14}$$

From (2.14), it follows the estimate

$$\big|s(\Phi^{(k)} \backslash X_\alpha(\eta_k, t_0, x)) - s(\Phi^{(k)} \backslash \widetilde{X}^\Gamma_\beta(\eta_k, t_0, y))\big| \leqslant l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2, \quad \eta_k \in \mathcal{T}. \tag{2.15}$$

From (2.15), we obtain the estimate

$$\Big|\sum_{\eta_k \in \mathcal{T}} s(\Phi^{(k)} \backslash X_\alpha(\eta_k, t_0, x)) - \sum_{\eta_k \in \mathcal{T}} s(\Phi^{(k)} \backslash \widetilde{X}^\Gamma_\beta(\eta_k, t_0, y))\Big| \leqslant N_* \cdot \big(l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2\big),$$

which can be written in the form

$$|J^{(1)}(\alpha, x) - \widetilde{J}^{(1)}_\Gamma(\beta, y)| \leqslant N_* \cdot \big(l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2\big). \tag{2.16}$$

A similar scheme is used to derive the estimate

$$|J^{(2)}(\alpha, x) - \widetilde{J}^{(2)}_\Gamma(\beta, y)| \leqslant l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2. \tag{2.17}$$

From estimates (2.16) and (2.17), we obtain

$$|J(\alpha, x) - \widetilde{J}_\Gamma(\beta, y)| \leqslant \zeta^\Delta(\rho, \sigma), \tag{2.18}$$

where

$$\zeta^\Delta(\rho, \sigma) = (N_* + 1) \cdot \big(l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \cdot \varkappa^\Delta(\rho, \sigma)^2\big);$$

Δ, ρ, and σ are from $(0, \infty)$.

Based on esimate (2.18), we show that, for small Δ, ρ, and σ, the solutions of Problems 2 and 3 are close, and we estimate this proximity.

Indeed, according to (2.18), the following inequality holds for every pair $(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)} \subset \mathscr{L} \times X^{(0)}$:

$$|J(\beta, y) - \widetilde{J}_\Gamma(\beta, y)| \leqslant \zeta^\Delta(\rho, \sigma),$$

since the pair $(\beta, y) \in \mathscr{L} \times X^{(0)}$ is the closest pair in $\mathscr{L}^{(\rho)} \times X^{(\sigma)}$ to itself and, therefore, satisfies the inequalities $\|\beta - \beta\| \leqslant \rho$ and $\|y - y\| \leqslant \sigma$.

Hence, the following inequality holds for every pair $(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$:

$$\widetilde{J}_\Gamma(\beta, y) - \zeta^\Delta(\rho, \sigma) \leqslant J(\beta, y) \leqslant \max_{(\alpha, x) \in \mathscr{L} \times X^{(0)}} J(\alpha, x),$$

which implies

$$\max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(0)}} \widetilde{J}_\Gamma(\beta,y)-\zeta^\Delta(\rho,\sigma)\leqslant \max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(x,\alpha). \tag{2.19}$$

On the other hand, according to (2.18), the inequality

$$J(\alpha,x)\leqslant \widetilde{J}_\Gamma(\beta,y)+\zeta^\Delta(\rho,\sigma)$$

is true for every $(\alpha,x)\in\mathscr{L}\times X^{(0)}$ and $(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}$ such that $\|\alpha-\beta\|\leqslant\rho$ and $\|x-y\|\leqslant\sigma$.

Hence, for every pair $(\alpha,x)\in\mathscr{L}\times X^{(0)}$, the inequality

$$J(\alpha,x)\leqslant \max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}} \widetilde{J}_\Gamma(\beta,y)+\zeta^\Delta(\rho,\sigma)$$

holds, which, in turn, implies

$$\max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(\alpha,x)\leqslant \max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}} \widetilde{J}_\Gamma(\beta,y)+\zeta^\Delta(\rho,\sigma). \tag{2.20}$$

Inequalities (2.19) and (2.20) imply

$$\max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(0)}} \widetilde{J}_\Gamma(\beta,y)-\zeta^\Delta(\rho,\sigma)\leqslant \max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(\alpha,x)\leqslant \max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}} \widetilde{J}_\Gamma(\beta,y)+\zeta^\Delta(\rho,\sigma),$$

i.e., we have the estimate

$$\Big|\max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(\alpha,x)-\max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(0)}} \widetilde{J}_\Gamma(\beta,y)\Big|\leqslant\zeta^\Delta(\rho,\sigma).$$

Let us say a pair $(\beta^*,y^*)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}$ is the optimal in the Problem 3, i.e.,

$$\widetilde{J}_\Gamma(\beta^*,y^*)=\max_{(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}} \widetilde{J}_\Gamma(\beta,y).$$

Then we have the estimate

$$\Big|\max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(\alpha,x)-\widetilde{J}_\Gamma(\beta^*,y^*)\Big|\leqslant\zeta^\Delta(\rho,\sigma). \tag{2.21}$$

In addition, as shown above, the pair (β^*,y^*), like any pair $(\beta,y)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}$, satisfies the inequality

$$|\widetilde{J}_\Gamma(\beta^*,y^*)-J(\beta^*,y^*)|\leqslant\zeta^\Delta(\rho,\sigma). \tag{2.22}$$

From (2.21) and (2.22), we obtain

$$\Big|\max_{(\alpha,x)\in\mathscr{L}\times X^{(0)}} J(\alpha,x)-J(\beta^*,y^*)\Big|\leqslant 2\zeta^\Delta(\rho,\sigma). \tag{2.23}$$

Inequality (2.23) states that every optimal pair $(\beta^*,y^*)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}$ for Problem 3 is $2\zeta^\Delta(\rho,\sigma)$-optimal for Problem 2.

Taking into account the quadratic dependence of the function $\zeta^\Delta(\rho,\sigma)$ from the function $\varkappa^\Delta(\rho,\sigma)$ and the equality $\lim\limits_{\Delta\downarrow0,\rho\downarrow0,\sigma\downarrow0}\varkappa^\Delta(\rho,\sigma)=0$, we obtain $\lim\limits_{\Delta\downarrow0,\rho\downarrow0,\sigma\downarrow0}\zeta^\Delta(\rho,\sigma)=0$. Hence, for a predetermined $\varepsilon>0$, one can choose $\Delta=\Delta(\Gamma)$, ρ, and σ from $(0,\infty)$ so that the following inequality is true:

$$\zeta^\Delta(\rho,\sigma)\leqslant\varepsilon. \tag{2.24}$$

Using ρ and σ satisfying (2.24), we can find a pair $(\beta^*,y^*)\in\mathscr{L}^{(\rho)}\times X^{(\sigma)}$ optimal for Problem 3.

As a result, the following statement is true.

Theorem 3. *Assume that the control system (1.1) in $\mathbb{R}^2$ satisfies the conditions $\boldsymbol{A}$ and $\boldsymbol{B}$ and, together with the partition $\Gamma = \{t_0, t_1, ..., t_i, ..., t_N = \vartheta\}$ ($\Delta = \Delta(\Gamma) = N^{-1}(\vartheta - t_0)$), the condition $\boldsymbol{C}$. Assume that, in Problems 2 and 3, along with the conditions $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, the conditions $\boldsymbol{D}$ and $\boldsymbol{E}$ are satisfied for the sets $X^{(0)}$, $X^{(\vartheta)}$, $\Phi_k = \Phi(\eta_k)$, $X_\alpha(t_i, t_0, x)$, and $\widetilde{X}_\beta^\Gamma(t_i, t_0, y)$, where $(\alpha, x) \in \mathscr{L} \times X^{(0)}$, $(\beta, y) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$, $t_i \in \Gamma$, and $\eta_k \in \mathcal{T}$.*

Then every optimal pair $(\beta^*, y^*) \in \mathscr{L}^{(\rho)} \times X^{(\sigma)}$ in Problem 3 is a $2\zeta^{(\Delta)}(\rho, \sigma)$-optimal pair in Problem 2.

3. Example

In this section, we consider a nonlinear control system in $\mathbb{R}^2$ on the time interval $[t_0, \vartheta] = [0, 1]$ depending on parameter α:

$$\begin{cases} \dot{x}_1 = -x_2 \cdot \dfrac{1}{2}\Big(7 + \dfrac{1}{4}\cos(x_2) + \dfrac{1}{2}\sin(\alpha_1 t)\Big) + \hat{a}(x)\dfrac{\|x\|}{1 + \|x\|}u_1 + 0.1\alpha_1, \\ \dot{x}_2 = x_1 \cdot \dfrac{1}{2}\Big(7 + \dfrac{1}{4}\cos(x_1) + \dfrac{1}{2}\sin(\alpha_2 t)\Big) + \hat{a}(x)\dfrac{\|x\|}{1 + \|x\|}u_2 + 0.1\alpha_2, \\ x(0) \in X^{(0)}, \end{cases} \tag{3.1}$$

where

$$\hat{a}(x) = \begin{cases} 0.01 & \text{for} \quad \|x\| < 1, \\ \dfrac{0.01}{\|x\|} & \text{for} \quad \|x\| \geqslant 1, \end{cases}$$

$$\alpha = (\alpha_1, \alpha_2) \in I = \left[\frac{2}{3}, \frac{4}{3}\right] \times \left[-\frac{1}{3}, \frac{1}{3}\right],$$

$$u = (u_1, u_2) \in P = \{\overline{u} = (\overline{u}_1, \overline{u}_2) : \|\overline{u}\| \leqslant 1\}.$$

One problem of targeting the integral funnels of system (3.1) is formulated and solved in a soft setting close to Problem 3 from the previous section.

In this setting, the set $X^{(0)}$ of initial positions of system (3.1) is a closed set in $\mathbb{R}^2$ bounded by the Cassini oval

$$(x^2 + y^2)^2 - 2c^2(x^2 - y^2)^2 \leqslant a^4 - c^4,$$

where $a = 4.4$ and $c = 4$.

Along with $X^{(0)}$ in $\mathbb{R}^2$, the following two sets are also given:

(1) the rectangle $\Phi(t), t \in [0, 1]$, with the initial set $\Phi(t_0) = [-9, -3] \times [-10, -6]$ rotating in one direction around its center $(-6, -8)$ over time 0.01 at the angle $1°$;

(2) the ellipse

$$X^{(\vartheta)} = \left\{(x_1, x_2) \in \mathbb{R}^2 : \Big(\frac{2}{5}\Big)^2 (x_1 - 12)^2 + \Big(\frac{2}{15}\Big)^2 (x_2 + 5)^2 \leqslant 1\right\}.$$

There are given the rectangle $\Phi(t), t \in [0, 1]$, changing its orientation over time t, that we treat as a zone of dangerous stay during the entire period of time $[0, 1]$, and the ellipse $X^{(\vartheta)}$ that we treat as a target set for the system (3.1) at the terminal time $\vartheta = 1$.

We have the following two goals for the control system (3.1):

(1) assuming that the reachable sets $X_\alpha(t, t_0, X^{(0)}), \alpha \in \mathscr{L}, t \in [t_0, \vartheta]$, can intersect the sets $\Phi(t)$, we must strive to ensure that the total intersection area $\Phi(\eta_i) \bigcap X_\alpha(\eta_i, t_0, X^{(0)}), \eta_i \in \mathscr{T}$, will be as small as possible; here $\mathscr{T}$ is some finite set in $[t_0, \vartheta]$;

(2) we must strive to ensure that the area of the intersection $X^{(\vartheta)} \bigcap X_\alpha(\vartheta, t_0, X^{(0)})$ is as much as possible.

Let us formalize our targeting problem.

Introduce the notation

$$J^{(1)}(\alpha) = \sum_{\eta_k \in \mathscr{T}} s(\Phi(\eta_k) \backslash X_\alpha(\eta_k, t_0, X^{(0)})),$$
$$J^{(2)}(\alpha) = \underset{\alpha \in \mathscr{L}}{s}(X^{(\vartheta)} \bigcap X_\alpha(\vartheta, t_0, X^{(0)})),$$
$$J(\alpha) = \lambda_1 J^{(1)}(\alpha) + \lambda_2 J^{(2)}(\alpha),$$

where λ_1 and λ_2 are from [0,1], $\lambda_1 + \lambda_2 = 1$.

Problem 4 on targeting integral funnels of system (3.1) (soft setting). It is required to find $\alpha^* \in \mathscr{L}$ such that

$$J(\alpha^*) = \max_{\alpha \in \mathscr{L}} J(\alpha).$$

Remark 5. Problem 4 is close in setting to Problem 2 from Section 2 and differs from it by considering the sets $X_\alpha(t_k, t_0, X^{(0)})$, $\alpha \in \mathscr{L}$, with initial set $X^{(0)}$ instead of the sets $X_\alpha(t_k, t_0, x)$, $(\alpha, x) \in \mathscr{L} \times X^{(0)}$. This limitation will not affect the key estimates that we use in Problem 4.

Since we cannot solve Problem 4 exactly, we formulate an approximation problem in which, instead of the set $\mathscr{L}$, there is a ρ-net $\mathscr{L}^{(\rho)}$. The partition Γ is used as the interval $[t_0, \vartheta]$ and, instead of ideal reachable sets $X_\alpha(t, t_0, X^{(0)})$, their approximations $\tilde{X}^{\Gamma}_{\alpha^{(r)}}(t_i, t_0, X^{(0)})$, $\alpha^{(r)} \in \mathscr{L}^{(\rho)}$, $t_i \in \Gamma$, are used.

Let us introduce the notation

$$\tilde{J}^{(1)}_{\Gamma}(\alpha^{(r)}) = \sum_{\eta_i \in \mathscr{T}} s(\Phi(\eta_i) \backslash \tilde{X}^{\Gamma}_{\alpha^{(r)}}(\eta_i, t_0, X^{(0)})),$$
$$\tilde{J}^{(2)}_{\Gamma}(\alpha^{(r)}) = s(X^{(\vartheta)} \bigcap \tilde{X}^{\Gamma}_{\alpha^{(r)}}(\vartheta, t_0, X^{(0)}));$$

recall that $s(X)$ is the area of a set $X \subset \mathbb{R}^2$.

Let us clarify how we define the partition Γ in the approximation problem and the ρ-net $\mathscr{L}^{(\rho)}$. Assume that

$$\Gamma = \{t_0 = 0, t_1, ..., t_i, ..., t_N = \vartheta = 1\},$$

where $t_{i+1} - t_i = \Delta_i = \Delta = 0.01$ and $N = 100$;

$$\mathscr{L}^{(\rho)} = \mathscr{L}^{(\rho)}_1 \times \mathscr{L}^{(\rho)}_2,$$

where

$$\rho = \frac{\sqrt{2}}{15},$$
$$\mathscr{L}^{(\rho)}_1 = \left\{\alpha^{(l)}_1 \in \left[\frac{2}{3}, \frac{4}{3}\right], \ \alpha^{(0)}_1 = \frac{2}{3}, \ \alpha^{(l)}_1 = l^{(0)}_1 + \frac{l}{15}, \ l = \overline{1, 10}\right\},$$
$$\mathscr{L}^{(\rho)}_2 = \left\{\alpha^{(k)}_2 \in \left[-\frac{1}{3}, \frac{1}{3}\right], \ \alpha^{(0)}_2 = -\frac{1}{3}, \ \alpha^{(k)}_2 = l^{(0)}_2 + \frac{k}{15}, \ k = \overline{1, 10}\right\};$$

in addition, points of the set $\mathscr{L}^{(\rho)}$ are parameterized by the parameter $r=\overline{1,121}$ and are denoted by $\alpha^{(r)}$.

Assume that

$$\tilde{J}_\Gamma(\alpha^{(r)})=\lambda_1\tilde{J}_\Gamma^{(1)}(\alpha^{(r)})+\lambda_2\tilde{J}_\Gamma^{(2)}(\alpha^{(r)}),$$

where λ_1 and λ_2 are defined above. Let us formulate an approximation problem.

Problem $4^{(a)}$. It is required to find $\alpha^{(r^*)}\in\mathscr{L}^{(\rho)}$ such that

$$\tilde{J}_\Gamma(\alpha^{(r^*)})=\max_{\alpha^{(r)}\in\mathscr{L}^{(\rho)}}\tilde{J}_\Gamma(\alpha^{(r)}).$$

It is important for us not only to calculate the optimal result $\tilde{J}_\Gamma(\alpha^{(r^*)})$ in Problem $4^{(a)}$, but also find out how accurately it approximates the optimal result $J(\alpha^*)$ in Problem 4. In other words, we are also interested in an upper estimate of the quantity $|J(\alpha^*)-\tilde{J}_\Gamma(\alpha^{(r^*)})|$. Note that this estimate is completely analogous to estimate (2.18) from Section 2 with the only difference that here $\sigma=0$.

Let us calculate the numerical characteristics in Problem $4^{(a)}$ involved in an estimate of type (2.18). Some of them will turn out to be quite significant in size. This is connected both with the dynamics of system (3.1) and with the roughness of the approximations Γ and $\mathscr{L}^{(\rho)}$ of the sets $[t_0,\vartheta]$ and $\mathscr{L}$.

The right-hand side of system (3.1) has the form

$$f_\alpha(t,x,u)=\begin{pmatrix}-\dfrac{x_2}{2}\cdot\Big(7+\dfrac14\cos(x_2)+\dfrac12\sin(\alpha_1 t)\Big)\\ \dfrac{x_1}{2}\cdot\Big(7+\dfrac14\cos(x_1)+\dfrac12\sin(\alpha_2 t)\Big)\end{pmatrix}+\hat{a}(x)\cdot\frac{\|x\|}{1+\|x\|}u+0.1\alpha.$$

Let us estimate from above the value $\|f_\alpha(t,x,u)\|$:

$$\begin{aligned}&\|f_\alpha(t,x,u)\|\\ &\leqslant\sqrt{\frac{x_2^2}{4}\Big(7+\frac14\cos(x_2)+\frac12\sin(\alpha_1 t)\Big)^2+\frac{x_1^2}{4}\Big(7+\frac14\cos(x_1)+\frac12\sin(\alpha_2 t)\Big)^2}\\ &\quad+\hat{a}(x)\frac{\|x\|}{1+\|x\|}\|u\|+0.1\|\alpha\|\\ &\leqslant\frac{31}{16}\|x\|+0.01\cdot1+0.1\sqrt{\Big(\frac43\Big)^2+\Big(\frac13\Big)^2}<1.9375\|x\|+0.1475,\end{aligned}$$

i.e.,

$$\|f_\alpha(t,x,u)\|\leqslant1.9375\|x\|+0.1475. \tag{3.2}$$

Hence, under the condition **B** for system (3.1), we can set $\gamma=1.9375$.

Using the Cauchy–Bunyakovsky inequality and an inequality of the type $2ab\leqslant a^2+b^2$, we estimate the variation of $\|x(t)\|^2$ along the trajectory $x=x(t)$ of system (3.1):

$$\begin{aligned}\frac{d\|x\|^2}{dt}&=2\langle x,\dot{x}\rangle=2(x_1\dot{x}_1+x_2\dot{x}_2)\\ &=-x_1x_2\Big(7+\frac14\cos x_2+\frac12\sin(\alpha_1 t)\Big)+2x_1\hat{\alpha}(x)\frac{\|x\|}{1+\|x\|}u_1+0.2\alpha_1x_1\\ &\quad+x_1x_2\Big(7+\frac14\cos x_1+\frac12\sin(\alpha_2 t)\Big)+2x_2\hat{\alpha}(x)\frac{\|x\|}{1+\|x\|}u_2+0.2\alpha_2x_2\\ &=\frac{x_1x_2}{4}\big(\cos x_1-\cos x_2+2\sin(\alpha_2 t)-2\sin(\alpha_1 t)\big)+2(x_1u_1+x_2u_2)\hat{\alpha}(x)\frac{\|x\|}{1+\|x\|}+0.2(\alpha_1x_2+\alpha_2x_2)\end{aligned}$$

$$\leqslant \frac{6}{4}|x_1x_2| + 2\|x\|\cdot\|u\|\hat{a}(x)\frac{\|x\|}{1+\|x\|} + 0.2\|\alpha\|\cdot\|x\|$$
$$\leqslant \frac{3}{4}\|x\|^2 + 2\hat{\alpha}(x)\|x\| + 0.2\|x\|\sqrt{\left(\frac{4}{3}\right)^2 + \left(\frac{1}{3}\right)^2}$$
$$\leqslant \frac{3}{4}\|x\|^2 + 0.295\|x\| \leqslant \frac{3}{4}\|x\|^2 + \frac{3}{10}\|x\|.$$

Given the equality
$$\frac{d\|x\|^2}{dt} = 2\|x\|\frac{d\|x\|}{dt},$$
from the inequality
$$\frac{d\|x\|^2}{dt} \leqslant \frac{3}{4}\|x\|^2 + \frac{3}{10}\|x\|,$$
we get
$$\frac{d\|x\|}{dt} \leqslant \frac{3}{8}\|x\| + \frac{3}{20} \quad \text{for} \quad \|x\| \neq 0.$$

From this inequality, we easily deduce the estimate
$$\|x(t)\| \leqslant \|x(t_0)\|\cdot e^{3/8\cdot(t-t_0)} + \frac{3\cdot 8}{20\cdot 3}(e^{3/8\cdot(t-t_0)} - 1), \quad t \in [t_0, \vartheta] = [0, 1].$$

Hence,
$$\max_{t\in[t_0,\vartheta]} \|x(t)\| \leqslant \|x(t_0)\|e^{3/8} + \frac{2}{5}\cdot(e^{3/8} - 1). \tag{3.3}$$

Taking into account the equation for the Cassini oval, we obtain
$$\max_{x(t_0)\in X^{(0)}} \|x(t_0)\| = \sqrt{a^2 + c^2} = \sqrt{4^2 + 4.4^2} \approx 5.946. \tag{3.4}$$

From (3.3) and (3.4), we conclude that the following estimate holds for the trajectories $x(t)$, $x^{(0)} \in X^{(0)}$ of system (3.1):
$$\max_{t\in[t_0,\vartheta]} \|x(t)\| \leqslant 8.833.$$

As constraints and a closed domain D in the space of positions (t, x) containing all possible motions $(t, x(t))$ of the control system together with some of their neighborhood (with respect to the phase variable), we can take the set
$$D = \{(t, x) : t \in [t_0, \vartheta],\ \|x\| \leqslant 8.833 + \varepsilon\},$$
where we set $\varepsilon = 0.167$. In this case, we find that
$$D = \{(t, x) : t \in [0, 1],\ \|x\| \leqslant 9\}.$$

Estimate (3.2) and inclusion $(t, x) \in D$ imply
$$\|f_\alpha(t, x, u)\| \leqslant 1.9375\cdot\|x\| + 0.1475 \leqslant 1.9375\cdot 9 + 0.1475 = 17.585$$
for $\alpha \in \mathscr{L}$ and $u \in P$.

It follows the inequality
$$\max\left\{\|x(t)\| : t \in [t_0, \vartheta],\ x(t_0) \in X^{(0)}\right\} \leqslant 8.833.$$

The following inequality is valid for (t,x,u) and (t,y,u) from $D\times P$ and $\alpha\in\mathscr{L}$:

$$\begin{gathered}\|f_\alpha(t,x,u)-f_\alpha(t,y,u)\| \\ \leqslant \left\|\begin{pmatrix}(y_2-x_2)\cdot\dfrac{1}{2}\Big(7+\dfrac{1}{2}\sin(\alpha_1 t)\Big)\\ (x_1-y_1)\cdot\dfrac{1}{2}\Big(7+\dfrac{1}{2}\sin(\alpha_2 t)\Big)\end{pmatrix}\right\| + \left\|\begin{pmatrix}\dfrac{1}{8}\big(y_2\cos y_2-x_2\cos x_2\big)\\ \dfrac{1}{8}\big(x_1\cos x_1-y_1\cos y_1\big)\end{pmatrix}\right\| \\ +\left|\hat\alpha(x)\cdot\frac{\|x\|}{1+\|x\|}-\hat\alpha(y)\cdot\frac{\|y\|}{1+\|y\|}\right|\cdot\|u\|.\end{gathered}$$

Let us estimate from above each of the three terms on the right-hand side of this inequality. We have

$$\left\|\begin{pmatrix}(y_2-x_2)\cdot\dfrac{1}{2}\Big(7+\dfrac{1}{2}\sin(\alpha_1 t)\Big)\\ (x_1-y_1)\cdot\dfrac{1}{2}\Big(7+\dfrac{1}{2}\sin(\alpha_2 t)\Big)\end{pmatrix}\right\| \leqslant \varphi(t)\|x-y\|,$$

where

$$\varphi(t)=\begin{cases}\dfrac{1}{2}+\dfrac{1}{2}\sin\Big(\dfrac{4}{3}t\Big) & \text{for}\quad t<\dfrac{3\pi}{8},\\ 1 & \text{for}\quad t\geqslant\dfrac{3\pi}{8}.\end{cases}$$

Further, taking into account the inequality

$$\begin{gathered}\left\|\frac{d(x_k\cos(x_k))}{dx_k}\right\| = \|\cos(x_k)-x_k\sin(x_k)\| \leqslant \sqrt{1+x_k^2}\leqslant\sqrt{1+9^2}\approx 9.055,\\ (t,x_k)\in D,\quad k=1,2,\end{gathered}$$

we get

$$\left\|\begin{pmatrix}\dfrac{1}{8}\big(y_2\cos y_2-x_2\cos x_2\big)\\ \dfrac{1}{8}\big(x_1\cos x_1-y_1\cos y_1\big)\end{pmatrix}\right\| \leqslant \frac{9.055}{8}\|x-y\|.$$

Let us now estimate the third term. To do this, we introduce

$$R=\hat\alpha(x)\cdot\frac{\|x\|}{1+\|x\|}-\hat\alpha(y)\cdot\frac{\|y\|}{1+\|y\|}$$

and estimate $|R|$:

$$|R|=\left|\frac{\hat\alpha(x)\|x\|(1+\|y\|)-\hat\alpha(y)\|y\|(1+\|x\|)}{(1+\|x\|)\cdot(1+\|y\|)}\right| \leqslant \big|\hat\alpha(x)\|x\|-\hat\alpha(y)\|y\|+\|x\|\|y\|(\hat\alpha(x)-\hat\alpha(y))\big|.$$

We consider four cases for further estimation of $|R|$.

Case 1. $\|x\|<1$ and $\|y\|<1$. Then

$$\begin{gathered}\hat\alpha(x)=\hat\alpha(y)=0.01,\\ |R|\leqslant 0.01\big|\|x\|-\|y\|\big|.\end{gathered}$$

Case 2. $\|x\|\geqslant 1$ and $\|y\|\geqslant 1$. Then

$$\begin{gathered}\hat\alpha(x)=\frac{0.01}{\|x\|},\qquad \hat\alpha(y)=\frac{0.01}{\|y\|},\\ |R|\leqslant\big|0.01(\|y\|-\|x\|)\big| = 0.01\big|\|x\|-\|y\|\big|.\end{gathered}$$

Case 3. $\|x\| < 1$ and $\|y\| \geqslant 1$. Then

$$\hat{\alpha}(x) = 0.01, \quad \hat{\alpha}(y) = \frac{0.01}{\|y\|},$$

$$|R| \leqslant \left|0.01\|x\| - \frac{0.01}{\|y\|} \cdot \|y\| + \|x\| \cdot \|y\|\Big(0.01 - \frac{0.01}{\|y\|}\Big)\right| = 0.01 \cdot \big|1 - \|x\| \cdot \|y\|\big|.$$

Consider two subcases.

(a) $\|x\| \cdot \|y\| \geqslant 1$. Then

$$|R| \leqslant 0.01\big|1 - \|x\| \cdot \|y\|\big| \leqslant 0.01\big|1 - \|y\|\big| \leqslant 0.01\big|\|x\| - \|y\|\big| < 0.01\|x - y\|.$$

(b) $\|x\| \cdot \|y\| < 1$. Then

$$|R| \leqslant 0.01\big|1 - \|x\|\|y\|\big| \leqslant 0.01\big|1 - \|x\|\big| \leqslant 0.01\big|\|y\| - \|x\|\big| \leqslant 0.01\|y - x\|.$$

Case 4. $\|x\| \geqslant 1$ and $\|y\| < 1$. Since this case is similar to Case 3, we have the inequality $|R| \leqslant 0.01\|x - y\|$.

Thus, in all cases, we have the inequality $|R| \leqslant 0.01\|x - y\|$.

Taking this inequality into account, we obtain an estimate for the third term:

$$|R| \cdot \|u\| \leqslant 0.01 \cdot \|x - y\|.$$

As a result, for (t, x, u) and (t, y, u) from $D \times P$ and $\alpha \in \mathscr{L}$, we get

$$\|f_\alpha(t, x, u) - f_\alpha(t, y, u)\| \leqslant L(t)\|x - y\|,$$

where $L(t) = \varphi(t) + 1.142$, and we can take $L = 2.142$.

Let us now estimate from above the value

$$\|f_\alpha(t, x, u) - f_\beta(\tau, x, u)\|,$$

where (t, x, u) and (τ, x, u) from $D \times P$ and α and β are from $\mathscr{L}$:

$$\begin{aligned}
\|f_\alpha(t, x, u) - f_\beta(\tau, x, u)\| &\leqslant \left\| \begin{pmatrix} -\frac{1}{4}x_2 \cdot (\sin(\alpha_1 t) - \sin(\beta_1 \tau)) \\ \frac{1}{4}x_1 \cdot (\sin(\alpha_2 t) - \sin(\beta_2 \tau)) \end{pmatrix} \right\| + 0.1 \left\| \begin{pmatrix} \alpha_1 - \beta_1 \\ \alpha_2 - \beta_2 \end{pmatrix} \right\| \\
&= \frac{1}{4}\sqrt{x_2^2 \cdot (\sin(\alpha_1 t) - \sin(\beta_1 \tau))^2 + x_1^2 \cdot (\sin(\alpha_2 t) - \sin(\beta_2 \tau))^2} + 0.1\|\alpha - \beta\| \\
&\leqslant \frac{1}{4}\|x\|\sqrt{(\alpha_1 t - \beta_1 \tau)^2 + (\alpha_2 t - \beta_2 \tau)^2} + 0.1\|\alpha - \beta\| \\
&\leqslant \frac{1}{4}\|x\|\big(|\alpha_1 t - \beta_1 \tau| + |\alpha_2 t - \beta_2 \tau|\big) + 0.1\|\alpha - \beta\| \\
&= \frac{1}{4}\|x\|\big(|\alpha_1 t - \alpha_1 \tau + \alpha_1 \tau - \beta_1 \tau| + |\alpha_2 t - \alpha_2 \tau + \alpha_2 \tau - \beta_2 \tau|\big) + 0.1\|\alpha - \beta\| \\
&\leqslant \frac{1}{4}\|x\|\big(\alpha_1|t - \tau| + \tau|\alpha_1 - \beta_1| + \alpha_2|t - \tau| + \tau|\alpha_2 - \beta_2|\big) + 0.1\|\alpha - \beta\| \\
&\leqslant \frac{1}{4}\|x\|\big((\alpha_1 + \alpha_2)|t - \tau| + \tau(|\alpha_1 - \beta_1| + |\alpha_2 - \beta_2|)\big) + 0.1\|\alpha - \beta\| \\
&\leqslant \frac{1}{4}\|x\|\Big(\frac{5}{3}|t - \tau| + \tau\sqrt{2}\|\alpha - \beta\|\Big) + 0.1\|\alpha - \beta\|
\end{aligned}$$

$$\leqslant \frac{1}{4}\max_{(t,x)\in D}\|x\|\cdot\big(\frac{5}{3}|t-\tau|+\sqrt{2}\|\alpha-\beta\|\big)+0.1\|\alpha-\beta\|$$
$$\leqslant \frac{1}{4}\max_{(t,x)\in D}\|x\|\cdot\max\Big(\frac{5}{3},\sqrt{2}+\frac{0.2}{\max_{(t,x)\in D}\|x\|}\Big)\cdot(|t-\tau|+\|\alpha-\beta\|)$$
$$=\frac{1}{4}\cdot 9\cdot\frac{5}{3}\cdot(|t-\tau|+\|\alpha-\beta\|)=\frac{15}{4}\cdot(|t-\tau|+\|\alpha-\beta\|).$$

As a result, we obtain the following estimate for (t,x,u) and (τ,x,u) from $D\times P$ and α and β from $\mathscr{L}$:

$$\|f_\alpha(t,x,u)-f_\beta(\tau,x,u)\|\leqslant\frac{15}{4}(|t-\tau|+\|\alpha-\beta\|),$$

from which it follows that, in the problem under consideration, we can take

$$\omega^*(\xi)=\frac{15}{4}\xi,\quad \xi\in(0,\infty).$$

So, we have calculated the main characteristics involved in this problem in an estimate of the type of estimate (2.18): $K, L(t), t\in[0,1], L\in(0,\infty)$, and $\omega^*(\xi),\xi\in(0,\infty)$. Let us supplement them with several more characteristics participating in this estimate. Namely, the performed calculations show that the lengths of the boundaries of the sets $X^{(0)}$, $X^{(\vartheta)}$, $\Phi(t_i)$, $t_i\in\Gamma$, and $\tilde{X}^{\Gamma}_{\alpha^{(r)}}(t_i,t_0,X^{(0)}), t_i\in\Gamma,\alpha^{(r)}\in\mathscr{L}^{(\rho)}$, can be etsimated by the number $l^*=32$. We also assume that $\Delta=\Delta(\Gamma)=\Delta_i=t_{i+1}-t_i=1/N=0.01$, where $N=100$, $N_*=N$, $\rho=1/15$, and $\sigma=0$ as noted above.

Having determined all the main numerical characteristics in the considered problem, we proceed to calculating the main estimate of the type of estimate (2.18).

The following relations are valid:

$$\varkappa^\Delta(\rho,\sigma)=e^{\sum_{k=0}^{N-1}L(t_k)\Delta_k}\cdot\Big((\vartheta-t_0)\cdot\omega^*(\rho)+2(\vartheta-t_0)\omega^*(\Delta)+2K\sum_{k=0}^{N-1}L(t_k)\Delta_k+\sigma\Big)$$
$$\approx e^{1.926}\cdot\Big(\frac{15}{4}\rho+2\cdot\frac{15}{4}\Delta+2\cdot 17.585\Delta\cdot 1.926\Big)\approx 516.3\cdot\Delta+25.73\cdot\rho;$$
$$\zeta^\Delta(\rho,\sigma)=(N_*+1)\cdot\big(l^*\cdot\varkappa^\Delta(\rho,\sigma)+\pi\varkappa^\Delta(\rho,\sigma)^2\big)$$
$$\approx(N+1)\cdot(32\cdot(516.3\cdot\Delta+25.73\cdot\rho)+3.142\cdot(516.3\cdot\Delta+25.73\cdot\rho)^2\big)\approx 37242.74.$$

Hence, we obtain the following estimate of the mismatch of the optimal results $J(\alpha^*)$ and $\tilde{J}_\Gamma(\alpha^*_r)$ in Problems 4 and $4^{(a)}$:

$$|J(\alpha^*)-\widetilde{J}_\Gamma(\alpha^{r^*})|\leqslant\zeta^\Delta(\rho,\sigma)\approx 37242.74. \tag{3.5}$$

Remark 6. Estimating the mismatch between the optimal results in Problems 4 and $4^{(a)}$, we found that estimate (3.5) is very rough. The roughness of estimate (3.5) is due to several factors:

(1) the dynamics of system (3.1);

(2) the presence of exponential quantities in the derivation of an estimate, which is standard for control problems with a Lipschitz right-hand side of the control system with a phase variable;

(3) the roughness of our discrete approximations $\mathscr{L}^{(\rho)}$ and Γ of the compact set $\mathscr{L}$ and the interval [0,1] due to the limited capabilities of computer technology.

Note that, although estimate (3.5) is rough, it was obtained within the framework of the theory developed in Sections 1 and 2 and does not reflect the real value of the mismatch $|J(\alpha^*) - \widetilde{J}_\Gamma(\alpha^{r^*})|$, which is much smaller.

Nevertheless, the question arises, how and by what means can estimate (3.5) be improved. One way to improve this is to establish more accurate approximations $\mathscr{L}^{(\rho)}$ and Γ for the compact set $\mathscr{L}$ and the time interval $[t_0, \vartheta] = [0, 1]$. In addition, if in the setting of Problems 4 and $4^{(a)}$, the number N_* is small, then it also improves estimate (3.5).

As an example, let us set $\rho = 1/150$, $N_* = 1$, $N = 1000$, and therefore $\Delta = \Delta(\Gamma) = 0.001$. Then

$$\begin{aligned}\varkappa^\Delta(\rho, \sigma) &= e^{\sum_{k=0}^{N-1} L(\eta_k)\Delta_k} \cdot \Big((\vartheta - t_0) \cdot \omega^*(\rho) + 2(\vartheta - t_0)\omega^*(\Delta) + 2K\Delta \sum_{k=0}^{N-1} L(\eta_k)\Delta_k\Big) \\ &\approx e^{1.928}\Big(\frac{15}{4}\rho + 2 \cdot \frac{15}{4}\Delta + 2 \cdot 17.585 \cdot \Delta \cdot 1.928\Big) \approx 517.8 \cdot \Delta + 25.78 \cdot \rho,\end{aligned}$$

$$\begin{aligned}\zeta^\Delta(\rho, \sigma) &= (N_* + 1) \cdot \big(l^* \cdot \varkappa^\Delta(\rho, \sigma) + \pi \varkappa^\Delta(\rho, \sigma)^2\big) \\ &\approx 2 \cdot \big(32 \cdot (517.8\Delta + 25.78\rho) + 3.142(517.8 \cdot \Delta + 25.78 \cdot \rho)^2\big) \approx 47.13.\end{aligned}$$

We see that the decrease in the values ρ and Δ by a factor of 10 and the number N_* by a factor of 100 led to a significant improvement in the value of $\zeta^\Delta(\rho, \sigma)$.

Note also that if the compact set $\mathscr{L}$ is finite by the statement of Problem $4^{(a)}$, then we can treat it as a finite approximation $\mathscr{L}^{(\rho)}$ of itself with the value $\rho = 0$.

In this case, with the same N_* and N as in the previous example, we get the estimate

$$|J(\alpha^*) - \widetilde{J}_\Gamma(\alpha^{r^*})| \leqslant \zeta^\Delta(\rho, \sigma) \approx 34.82.$$

For the example under consideration, we considered three variants of Problem 2 on targeting integral funnels (in a soft setting). Moreover, the peculiarity of our consideration is that we do not vary the starting point $x^{(0)}$ in the set $X^{(0)}$ and, instead of the sets $\widetilde{X}_\alpha^\Gamma(t_i, t_0, x^{(0)})$, consider the reachable sets $\widetilde{X}_\alpha^\Gamma(t_i, t_0, X^{(0)})$, $\alpha \in \mathscr{L}$.

Each of the variants is determined by the choice of a pair of numbers λ_1 and λ_2 ($\lambda_1, \lambda_2 \in [0, 1]$, $\lambda_1 + \lambda_2 = 1$):

Variant 1. $\lambda_1 = 0.1$ and $\lambda_2 = 0.9$;
Variant 2. $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$;
Variant 3. $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$.

For each of the variants, in the set $\mathscr{L}^{(\rho)} \subset \mathscr{L}$, the optimal point $\alpha^* = (\alpha_1^*, \alpha_2^*)$ is calculated:

Variant 1. $\alpha_1^* = 0.66667$ and $\alpha_2^* = 0.33333$;
Variant 2. $\alpha_1^* = 0.80000$ and $\alpha_2^* = 0.33333$;
Variant 3. $\alpha_1^* = 1.26670$ and $\alpha_2^* = 0.33333$.

Each of the three options is illustrated with six figures (Fig. 9–Fig. 14, Fig. 15–Fig. 20, Fig. 21–Fig. 26) that correspond to the times $t_i = 0$; 0.2; 0.4; 0.6; 0.8; 1.0 of the partition Γ. The figures show the sets $\widetilde{X}_{\alpha^*}^\Gamma(t_i, t_0, X^{(0)})$ and $\Phi(t_i)$ corresponding to these numbers and the target set M.

Also, each of the three variants indicates the optimal result obtained in the course of an approximate solution of Problem 2:

Variant 1. $J(\alpha^*, X^{(0)}) = 38.4361$;
Variant 2. $J(\alpha^*, X^{(0)}) = 743.9625$;
Variant 3. $J(\alpha^*, X^{(0)}) = 2450$.

Variant 1

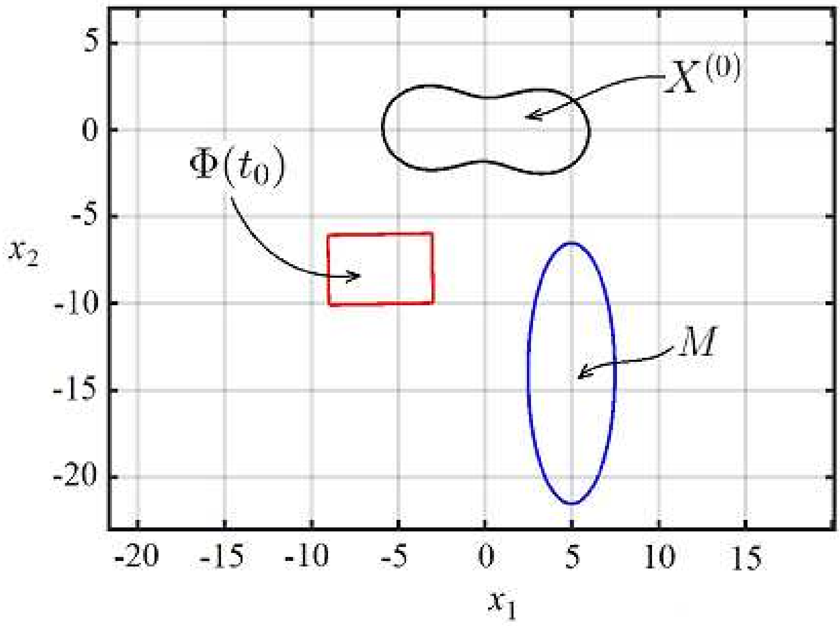


Figure 9. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.66667$, $\alpha_2^* = 0.33333$, and $t_i = 0$.

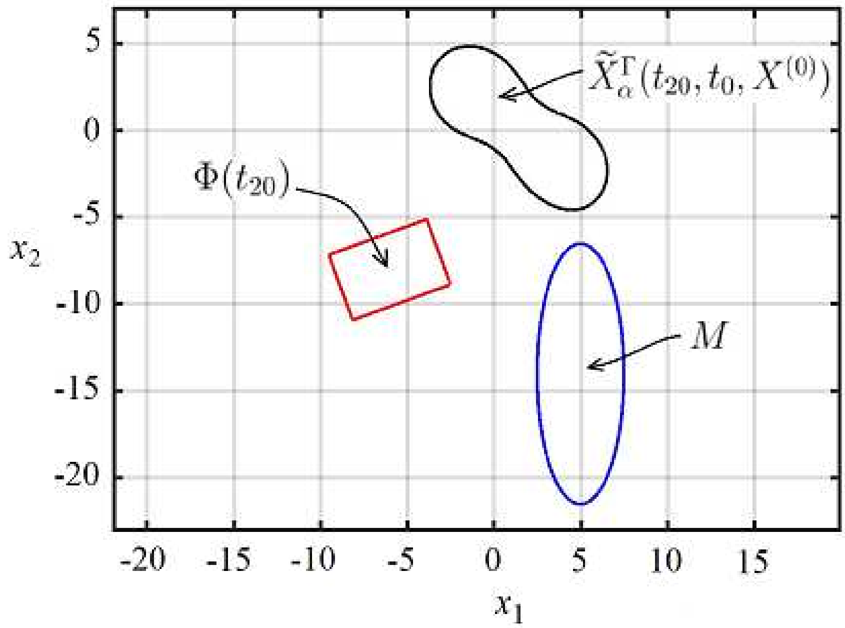


Figure 10. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.66667$, $\alpha_2^* = 0.33333$, and $t_i = 0.2$.

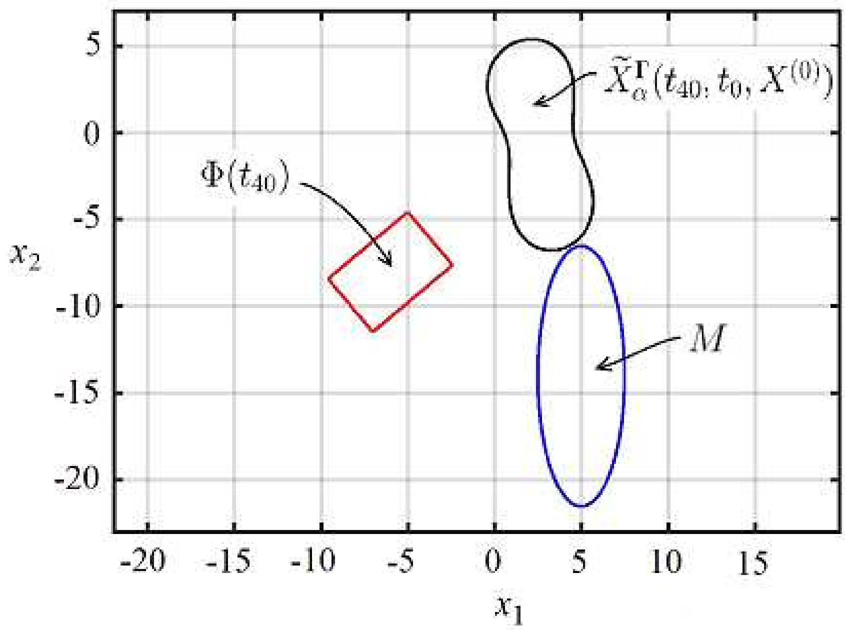


Figure 11. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.66667$, $\alpha_2^* = 0.33333$, and $t_i = 0.4$.

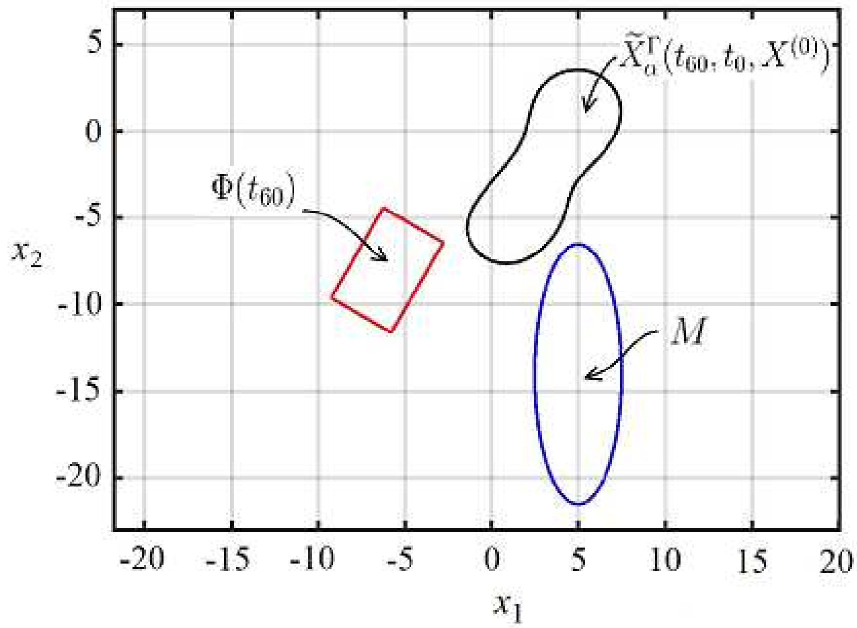


Figure 12. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.66667$, $\alpha_2^* = 0.33333$, and $t_i = 0.6$.

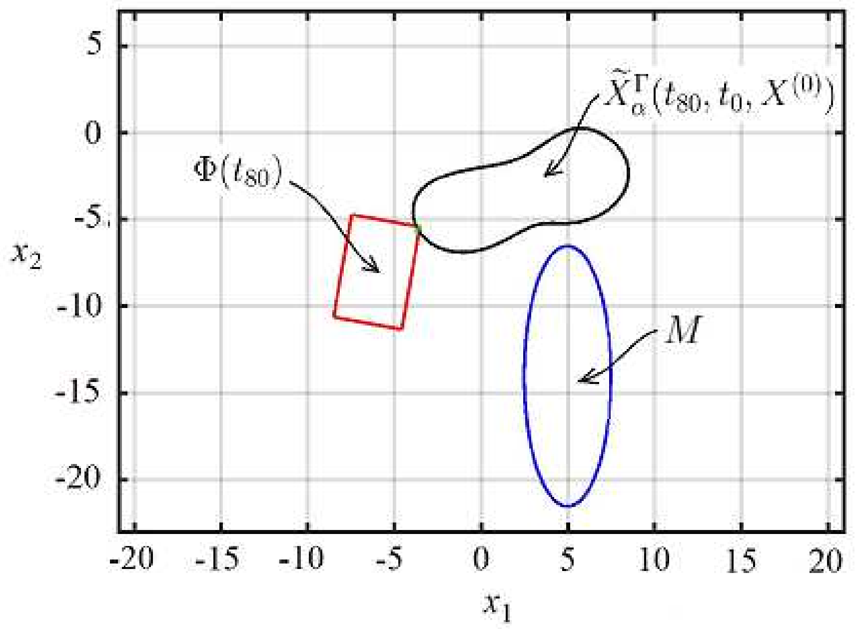


Figure 13. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.66667$, $\alpha_2^* = 0.33333$, and $t_i = 0.8$.

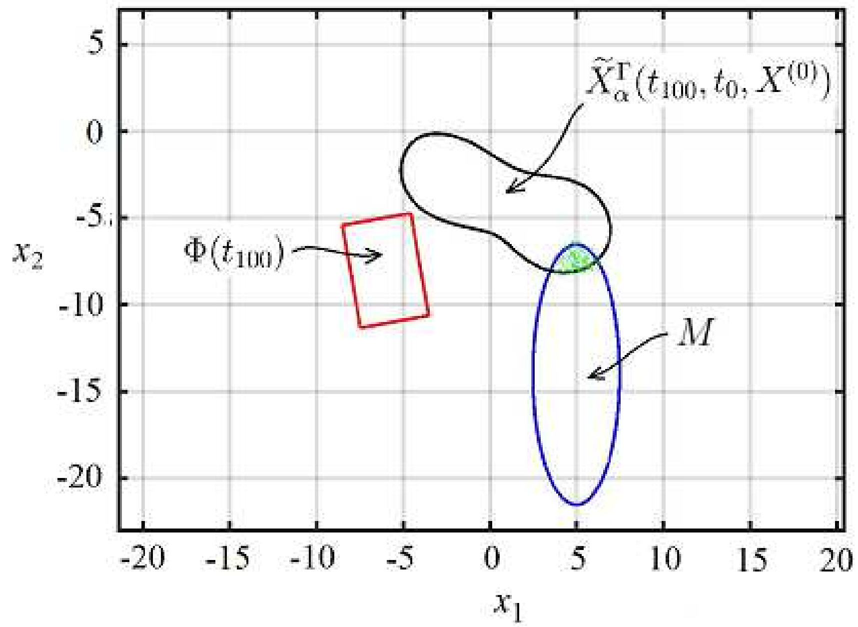


Figure 14. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.66667$, $\alpha_2^* = 0.33333$, and $t_i = 1$.

Variant 2

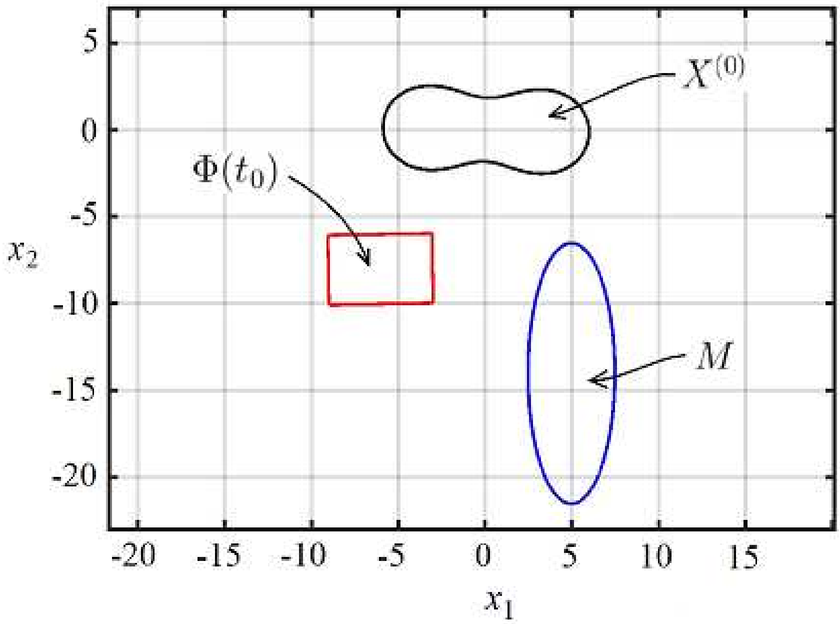


Figure 15. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.8$, $\alpha_2^* = 0.33333$, and $t_i = 0$.

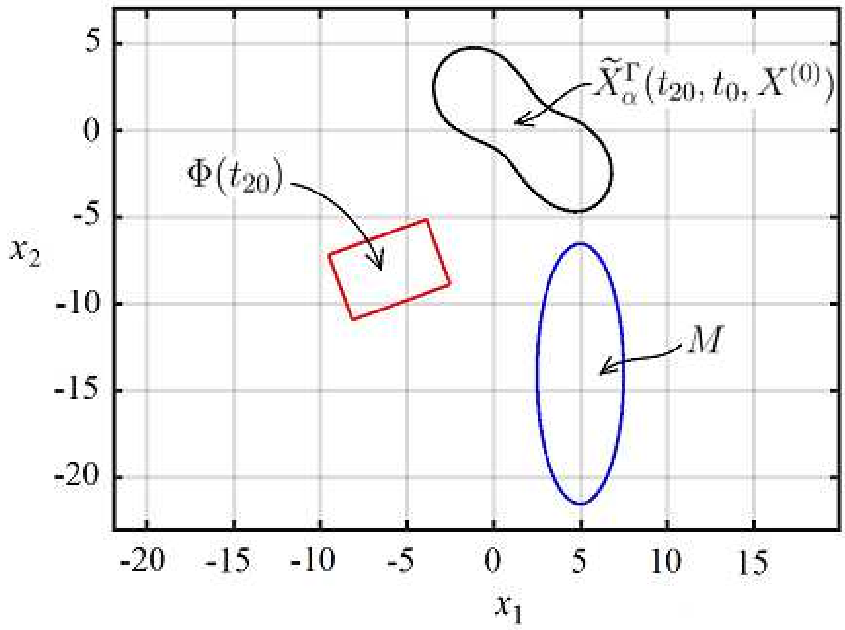


Figure 16. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.8$, $\alpha_2^* = 0.33333$, and $t_i = 0.2$.

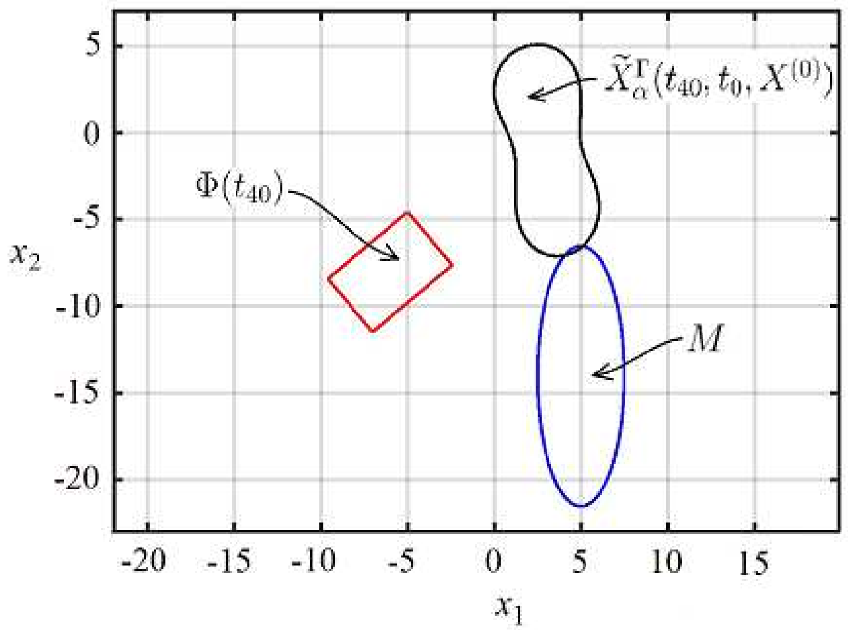


Figure 17. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.8$, $\alpha_2^* = 0.33333$, and $t_i = 0.4$.

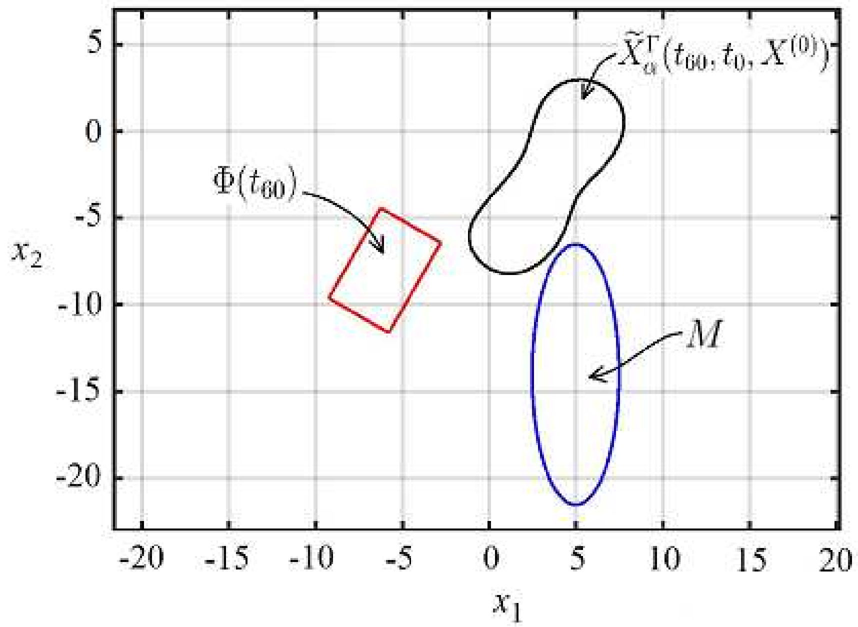


Figure 18. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.8$, $\alpha_2^* = 0.33333$, and $t_i = 0.6$.

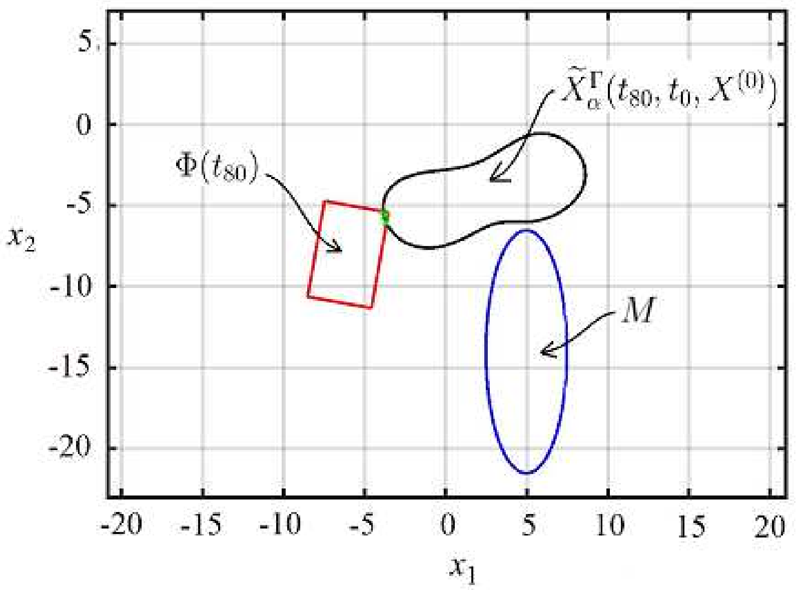


Figure 19. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.8$, $\alpha_2^* = 0.33333$, and $t_i = 0.8$.

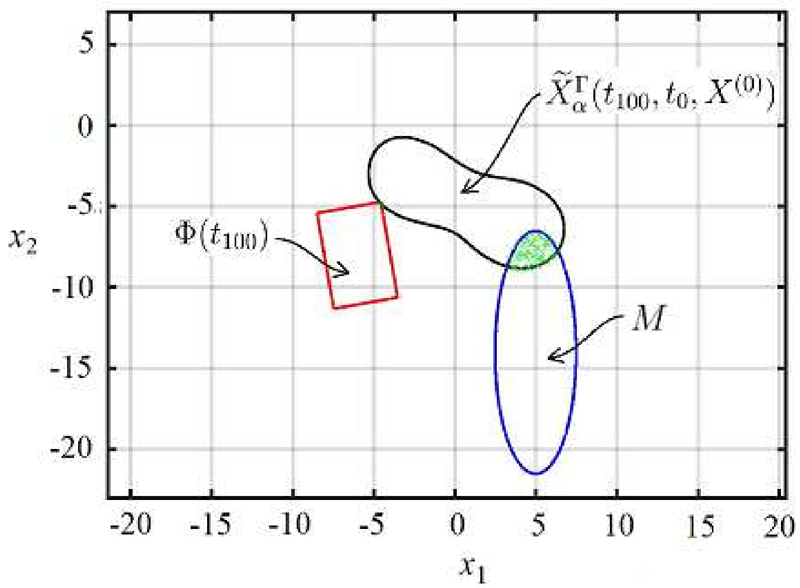


Figure 20. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 0.8$, $\alpha_2^* = 0.33333$, and $t_i = 1$.

Variant 3

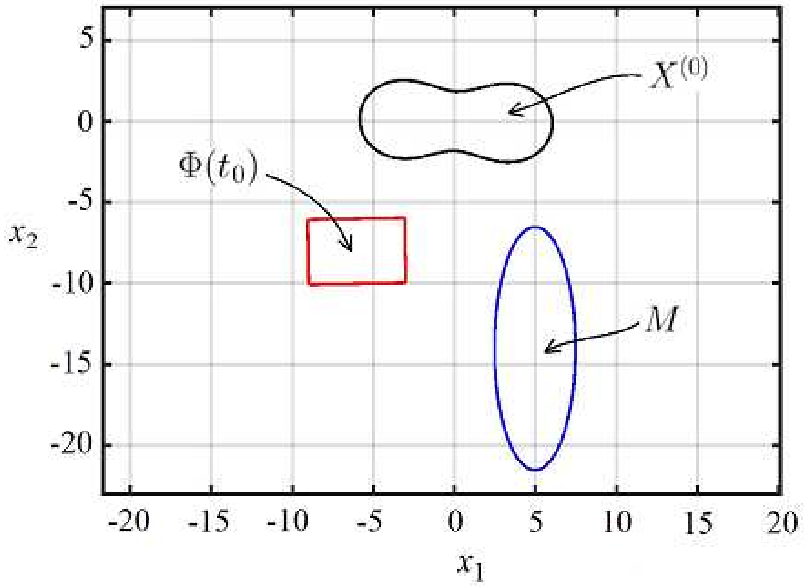


Figure 21. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 1.2667$, $\alpha_2^* = 0.33333$, and $t_i = 0$.

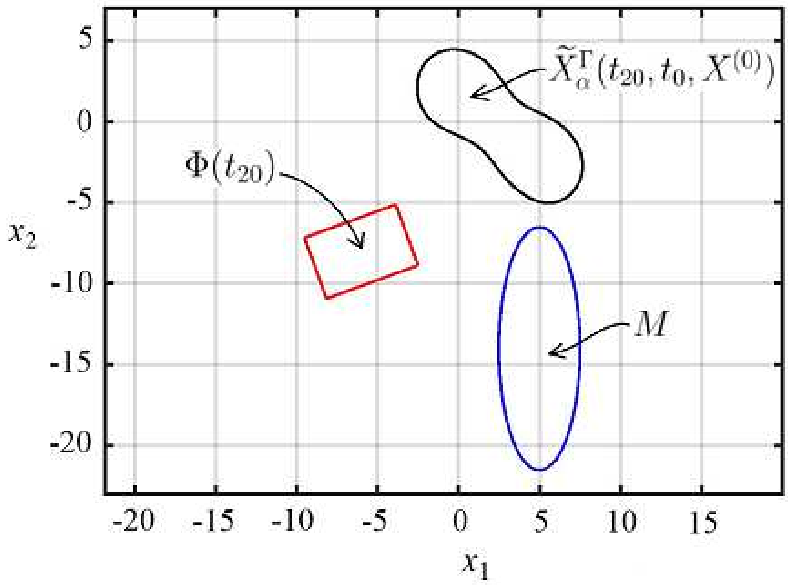


Figure 22. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 1.2667$, $\alpha_2^* = 0.33333$, and $t_i = 0.2$.

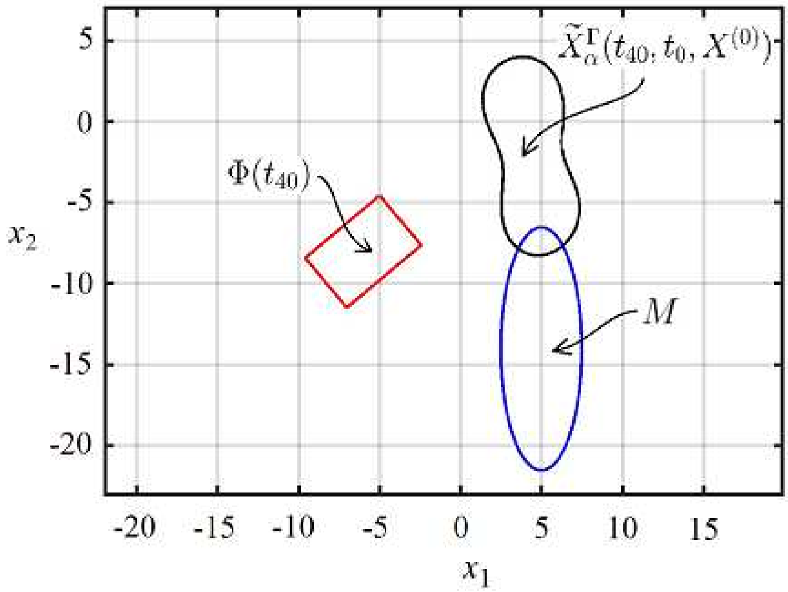


Figure 23. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 1.2667$, $\alpha_2^* = 0.33333$, and $t_i = 0.4$.

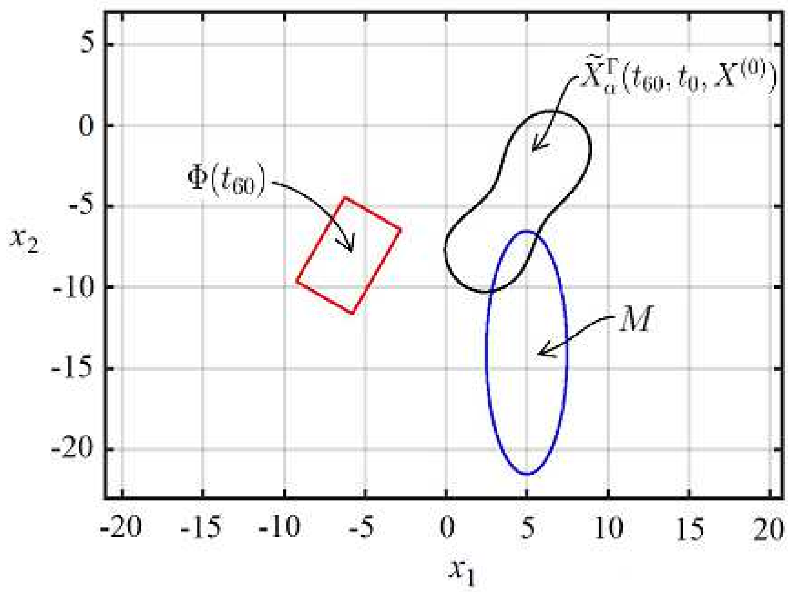


Figure 24. The sets M, $\Phi(t_i)$, and $\widetilde{X}^{\Gamma}_{\alpha}(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 1.2667$, $\alpha_2^* = 0.33333$, and $t_i = 0.6$.

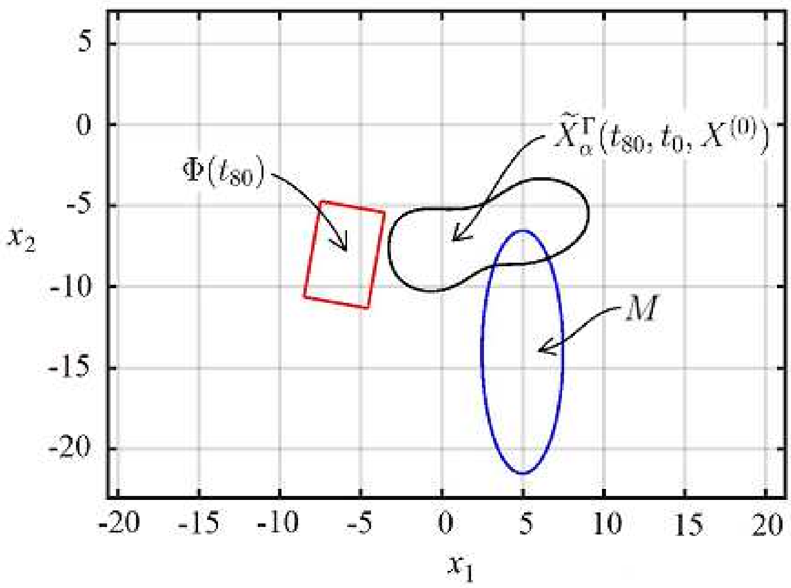


Figure 25. The sets M, $\Phi(t_i)$, and $\widetilde{X}_\alpha^\Gamma(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 1.2667$, $\alpha_2^* = 0.33333$, and $t_i = 0.8$.

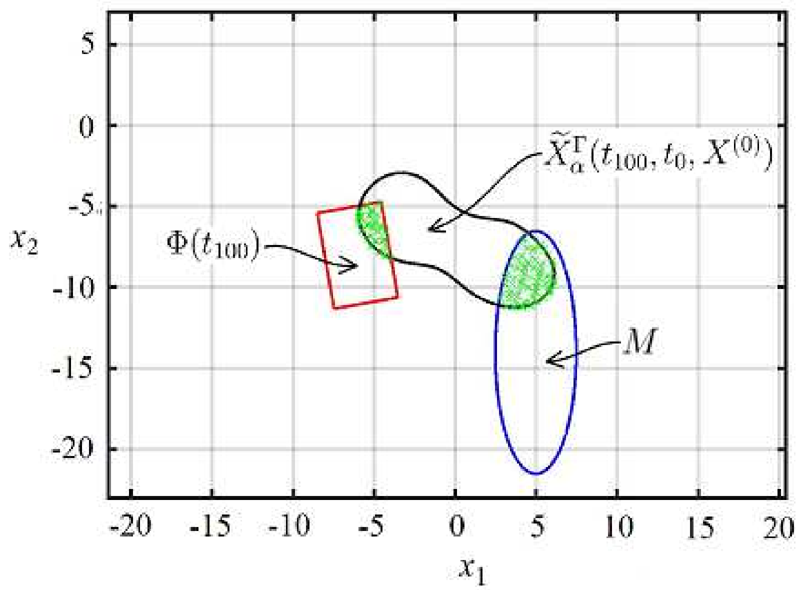


Figure 26. The sets M, $\Phi(t_i)$, and $\widetilde{X}_\alpha^\Gamma(t_i, t_0, X^{(0)})$ for $\alpha_1^* = 1.2667$, $\alpha_2^* = 0.33333$, and $t_i = 1$.

REFERENCES

1. Anan'evskii I. M. Control of a nonlinear vibratory system of the fourth order with unknown parameters. *Autom. Remote Control,* 2001. Vol. 62, No. 3. P. 343–355. DOI: 10.1023/A:1002832924913
2. Anan'evskii I. M. Control synthesis for linear systems by methods of stability theory of motion. *Differential Equations,* 2003. Vol. 39, No. 1. P. 1–10. DOI: 10.1023/A:1025170521270
3. Beznos A. V., Grishin A. A., Lensky A. V., Okhotsimsky D. E., Formalsky A. M. Pendulum control using a flywheel. In: *Spetspraktikum po teoreticheskoi i prikladnoi mehanike [Special workshop on theoretical and applied mechanics]. V. V. Aleksandrov, Yu. V. Bolotov (eds.).* Moscow: MSU Press, 2019. P. 170–195.
4. Bogachev V.I., Smoljanov O.G. *Deistvitel'nyi i funktsional'nyi analiz: universitetskii kurs* [Real and Functional Analysis: University Course]. Moscow-Izhevsk: Research Center "Regular and Chaotic Dynamics", Institute for Computer Research, 2009. 724 p. (in Russian)
5. Chernousko F. L. *State Estimation for Dynamic Systems.* CRC Press: Boca Raton, 1994. 320 p.
6. Chernousko F. L., Melikyan A. A. *Igrovye zadachi upravlenija i poiska* [Game Control and Search Problems]. Moscow: Nauka, 1978. 270 p. (in Russian)
7. Ershov A. A., Ushakov V. N. An approach problem for a control system with an unknown parameter. *Sb. Math.,* 2017. Vol. 208. No. 9. P. 1312–1352. DOI: 10.1070/SM8761
8. Filippova T. F. Construction of set-valued estimates of reachable sets for some nonlinear dynamical systems with impulsive control. *Proc. Steklov Inst. Math.*, 2010. Vol. 269, Suppl. 1. P. S95–S102. DOI: 10.1134/S008154381006009X
9. Gusev M. I. Estimates of reachable sets of multidimensional control systems with nonlinear interconnections. *Proc. Steklov Inst. Math.,* 2010. Vol. 269, Suppl. 1. P. S134–S146. DOI: 10.1134/S008154381006012X
10. Krasovsky N. N. *Upravlenie dinamicheskoi sistemoi: Zadacha o minimume garantirovannogo rezul'tata* [Control of a Dynamical System: Problem on the Minimum of Guaranteed Result]. Moscow: Nauka, 1985. 520 p. (in Russian)
11. Krasovsky N. N., Subbotin A. I. *Pozitsionnye differentsial'nye igry* [Positional Differential Games]. Moscow: Fizmatlit, 1974. 456 p. (in Russian)
12. Kurzhansky A. B. *Izbrannye trudy* [Selected Works]. Moscow: MSU Press, 2009. 756 p. (in Russian)
13. Kurzhanski A. B., Valyi I. *Ellipsoidal Calculus for Estimation and Control.* Systems Control Found. Appl. Basel: Birkhäuser, 1997. 321 p.
14. Lee E. B., Markus L. *Foundation of Optimal Control Theory.* New York–London–Sydney: John Wiley & Sons, 1967. 576 p.
15. Leichtweiß K. *Konvexe Mengen.* Hochschultext. Berlin: Springer-Verlag, 1979. 330 p. (in German)
16. Lempio F., Veliov V. M. Discrete approximation of differential inclusions. *Bayreuth. Math. Schr.*, 1998. Vol. 54. P. 149–232.
17. Nikol'skii M. S. On the approximation of the reachable set of a differential inclusion. *Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet*, 1987. No. 4. P. 31–34.
18. Nikol'skii M. S. An inner estimate of the attainability set of Brockett's nonlinear integrator. *Differential Equations,* 2000. Vol. 36, No. 11. P. 1647–1651. DOI: 10.1007/BF02757366
19. Polyak B. T., Khlebnikov M. V., Shcherbakov P. S. *Upravleniye lineynymi sistemami pri vneshnih vozmushcheniyah: Tehnika lineynyh matrichnyh neravenstv* [Control of linear systems under external disturbances: Technique of linear matrix inequalities]. Moscow: LENAND, 2014. 560 p. (in Russian)
20. Ushakov V. N., Matviychuk A. R., Ushakov A. V. Approximations of attainability sets and of integral funnels of differential inclusions. *Vestn. Udmurtsk. Univ. Mat. Mekh. Komp. Nauki*, 2011. No. 4. P. 23–39. (in Russian) URL: http://vst.ics.org.ru/journal/article/1816/
21. Vdovin S. A., Taras'yev A. M., Ushakov V. N. Construction of the attainability set of a Brockett integrator. *J. Appl. Math. Mech.,* 2004. Vol. 68, No. 5. P. 631–646. DOI: 10.1016/j.jappmathmech.2004.09.001

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 160–167

DOI: 10.15826/umj.2021.1.012

SET MEMBERSHIP ESTIMATION WITH A SEPARATE RESTRICTION ON INITIAL STATE AND DISTURBANCES[1]

Polina A. Yurovskikh

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya Str., Ekaterinburg, 620108 Russia

polina2104@list.ru

Abstract: We consider a set membership estimation problem for linear non-stationary systems for which initial states belong to a compact set and uncertain disturbances in an observation equation are integrally restricted. We prove that the exact information set of the system can be approximated by a set of external ellipsoids in the absence of disturbances in the dynamic equation. There are three examples of linear systems. Two examples illustrate the main theorem of the paper, the latter one shows the possibility of generalizing the theorem to the case with disturbances in the dynamic equation.

Keywords: Set membership estimation, Filtration, Approximation, Information set, Ellipsoid approach.

1. Introduction and notations

Set membership approaches to estimation problems have been studied for a long time [3, 10]. In 1968, Krasovskii proposed [7], and later Kurzhanski developed [8, 9] a more general theory of guaranteed estimation without the statistics of disturbances based on results of convex and functional analysis.

This paper is an addition to [2] which describes the approximation of the estimation problem for joint constraints on the initial state and disturbances with the ellipsoid technique. In this case, an optimization problem arises. The paper considers a simpler case when the exact information set of the system can be found without solving an optimization problem. The technique of ellipsoidal approximation is used, which was developed by Kurzhanski [9], Chernousko [4], and their followers (see, for example, [6]).

The paper is structured as follows. First, we formulate the estimation problem in our case, then construct exact information sets and their approximation using external ellipsoids. After that, we prove the validity of the approximation. The latter part consists of three numerical examples.

Let us introduce the notation. Let

$$|x|_Q = \sqrt{x'Qx},$$

where $x \in \mathbb{R}^n$ and Q is a matrix with the property $Q' = Q > 0$. For $Q = I$ (an identity matrix), we set $|x|_I = |x|$. If $M \subset \mathbb{R}^n$ is convex and compact, then

$$\rho(l|M) = \max_{x \in M} l'x$$

[1]This study is a part of the research carried out at the Ural Mathematical Center and supported by the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-02-2021-1383).

is a support function. The set

$$E(Q,c) = \{x \in \mathbb{R}^n \mid |x - c|_Q \leqslant 1\}$$

is called an ellipsoid.

If a system is linear and non-stationary, i.e., $\dot{x} = A(t)x$, then its general solution has the form $x(t, t_0, x_0) = \mathbf{X}(t, t_0)x_0$, where $\mathbf{X}(t, t_0)$ is a fundamental matrix, which can be found as a solution to the equation $\dot{\mathbf{X}}(t, t_0) = A(t)\mathbf{X}(t, t_0)$, $\mathbf{X}(t_0, t_0) = I$.

2. Problem statement

Consider a linear non-stationary system with measurements

$$\dot{x} = A(t)x, \quad y = G(t)x + w, \quad t \in [0, T], \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$ is a state vector, $y(t) \in \mathbb{R}^m$ is an output, $w(t) \in \mathbb{R}^m$ is an uncertain disturbance in the measurement equation, and $A(t) \in \mathbb{R}^{n\times n}$ and $G(t) \in \mathbb{R}^{m\times n}$ are bounded continuous matrices. Suppose that undefined functions $w(\cdot)$ in (2.1) and an initial state x_0 satisfy the following integral and geometric constraints, respectively:

$$\int_0^T |w(t)|_R^2 \leqslant 1, \quad R'(t) = R(t) > 0, \tag{2.2}$$

$$x_0 \in X_0, \tag{2.3}$$

where $X_0 \in \mathbb{R}^n$ is a convex compact set bounding the initial state, and $R(t) \in \mathbb{R}^{m\times m}$ is a continuous positive definite matrix. The constraints are separate, i.e., (2.2) and (2.3) are independent. According to the general theory of guaranteed estimation (see, for example, [9]) we can give a definition.

Definition 1. *A family of state vectors $\mathcal{X}(T, y) = \{x_T\}$ is called an information set (IS) if, for any $x_T \in \mathcal{X}(T, y)$, there exists a function w and an initial state x_0 satisfying constraints (2.2) and (2.3) and such that equalities (2.1) hold almost everywhere with $x(T) = x_T$.*

For system (2.1) under constraints (2.2) and (2.3), an exact set $\mathcal{X}(T, y)$ can be found.

Theorem 1. *The set $\mathcal{X}(T, y)$ is an intersection*

$$\mathcal{X}(T, y) = \mathbf{X}(T, 0)X_0 \bigcap \mathbb{X}(T, y, 0, 0),$$

where $\mathbf{X}(T, s)$ is the fundamental matrix of system (2.1), $\mathbb{X}(T, y, 0, 0)$ is the IS for (2.1) and (2.2) without constraints on the initial set (2.3).

Consider linear system (2.1) under constraint (2.2). A solution to the estimation problem is the IS $\mathbb{X}(T, y, 0, 0)$, which is an ellipsoid $x'P(T)x - 2x'd(T) + q(T) \leqslant 1$ whose parameters can be found as solutions to the differential equations [1, 2]

$$\dot{P}(t) = -A'(t)P - PA(t) + G'RG, \quad P(0) = 0;$$
$$\dot{d}(t) = -A'(t)d + G'Ry, \quad d(0) = 0;$$
$$\dot{q}(t) = y'Ry, \quad q(0) = 0.$$

3. Approximation of information sets

The original problem included integral constraints on perturbations (2.2) and geometric constraints on the initial state (2.3) of the system. Geometric constraints in form (2.3) are complicated to deal with. Kurzhanski proposed an approach for approximating arbitrary sets (see, for example, [9]) by sets of ellipsoids. In this paper, we discuss the approximation by a set of external ellipsoids.

We approximate the set of initial states X_0 by a family of ellipsoids $E(P_0, c) \supset X_0$, where P_0 is a symmetric positive definite matrix $P_0' = P_0 > 0$. Then constraints (2.2) and (2.3) will be approximated by the family of constraints

$$\alpha|x_0|^2_{P_0} + (1-\alpha)\int_0^T |w(t)|^2_R \leqslant 1, \quad \alpha \in [0,1]. \tag{3.1}$$

Thus, we obtain the second estimation problem of (2.1) under constraints (3.1).

If disturbances $w(t)$ satisfy the constraint in (2.2), then they necessarily obey the constraints in (3.1). Therefore, it is possible to build an IS $\mathbb{X}(T, y, \alpha, P_0)$ for a real signal with different parameters and use it to approximate the original IS $\mathcal{X}(T, y)$.

Lemma 1. *The set $\mathbb{X}(T, y, \alpha, P_0)$ has the form of an ellipsoid*

$$x'P(T)x - 2x'd(T) + q(T) \leqslant 1,$$

where the parameters are defined as solutions to the differential equations [1, 2]

$$\begin{gathered} \dot{P} = -A'(t)P - PA(t) + G'RG(1-\alpha), \quad P(0) = P_0\alpha; \\ \dot{d} = -A'(t)d + G'Ry(t)(1-\alpha), \quad d(0) = 0; \\ \dot{q} = y'(t)Ry(t)(1-\alpha), \quad q(0) = 0. \end{gathered} \tag{3.2}$$

Lemma 2 (Ellipsoid Separation Lemma). *For every convex compact set $M \subset \mathbb{R}^n$ and a point $p \notin M$, there exist an ellipsoid $E(Q, c)$ such that $E(Q, c) \supset M$ and $p \notin E(Q, c)$.*

P r o o f. It is known from convex analysis (see, for example, [5]), that the condition $p \notin M$ implies the existence of a unit vector l_1 such that $l_1'p > \rho(l_1|M)$. Further, since the set M is fixed, we use the shorter notation $\rho(l)$. Let us complement the vector l_1 to an orthonormal basis in $\mathbb{R}^n$ with vectors $\{l_2, \ldots, l_n\}$. Build a rectangular box along l_i centered at the point

$$c = \sum_{i=1}^n l_i(\rho(l_i) - \rho(-l_i))/2$$

and having vertices at the points

$$A_k = \sum_{i=1}^n k_i l_i \rho(k_i l_i) : \Pi = \left\{x \in \mathbb{R}^n \mid \rho(-l_i) \leqslant l_i'x \leqslant \rho(l_i) \ \forall i \in 1:n\right\}.$$

Here, $k \in K \subset \mathbb{R}^n$ is a vector with coordinates $k_i = \pm 1$. The number of such vectors and vertices is $2n$; the set K contains all such vectors k. Let us arrange the set $K = \left\{k^1, \ldots, k^{2n}\right\}$ assuming that $k^1 = [1; \ldots; 1]$. This box will contain the original compact set: $\Pi \supset M$.

We introduce an orthogonal matrix $T = [l_1, \ldots, l_n]$ and perform an orthogonal transformation to new coordinates $y = T'x$. In the new coordinates, the set M becomes $M^* = T'M$, and the box Π becomes the box $\Pi^* = T'\Pi$ with center $c^* = T'c$ and edges parallel to the coordinate axes. We have $l_1'p = l_1'Tp = [1, 0, \dots, 0]p^* = p_1^* > \rho(l_1)$ by the condition. We build an ellipsoid with the center c^* through the vertices of the box A_k^* and axes parallel to the coordinate axes, consisting of vectors of the form $c^* + y$, where the coordinates of the vector y satisfy the equation

$$\sum_{i=1}^{n} y_i^2/b_i^2 = 1. \tag{3.3}$$

Denote by $a_i = (\rho(l_i) + \rho(-l_i))/2$ the box semiaxes. Let us choose the parameters b_i of the ellipsoid so that

$$\sum_{i=1}^{n} a_i^2/b_i^2 = 1, \quad \rho(l_1) < b_1 + c_1^* < p_1^*. \tag{3.4}$$

Since $\rho(l_1) - c_1^* = a_1 < b_1$, the other b_i can be taken equal to $a_i + t$, $i \in 2:n$, where t can be found from the equation

$$\sum_{i=2}^{n} a_i^2/(a_i + t)^2 = 1 - a_1^2/b_1^2. \tag{3.5}$$

The obtained ellipsoid E^* with conditions (3.3), (3.4), and (3.5) is such that $E^* \supset \Pi^* \supset M^*$ and $p^* \notin E^*$. We obtain the desired ellipsoid with the properties $E(Q, c) \supset \Pi \supset M$ and $p \notin E(Q, c)$ by performing the inverse transformation $x = Ty$. Here, the matrix $Q = T\Lambda T'$ and $\Lambda = \mathrm{diag}(1/b_1, \dots, 1/b_n)$. □

Remark 1. If the set M is centrally symmetric, then $c = 0$.

Theorem 2. *Let X_0 be a centrally symmetric set. Then the set $\mathcal{X}(T, y)$ is an intersection*

$$\mathcal{X}(T, y) = \bigcap_{P_0 \in \mathcal{P}_0,\ \alpha} \mathbb{X}(T, y, P_0, \alpha),$$

where $\mathcal{P}_0$ is a set of symmetric positive matrices P_0 such that $E(P_0, 0) \supset X_0$.

P r o o f. We construct the proof by contradiction. Consider the inclusion

$$\mathcal{X}(T, y) \supset \bigcap_{P_0 \in \mathcal{P}_0, \alpha} \mathbb{X}(T, y, P_0, \alpha).$$

Let

$$x_* \in \bigcap_{P_0 \in \mathcal{P}_0, \alpha} \mathbb{X}(T, y, P_0, \alpha),$$

but $x_* \notin \mathcal{X}(T, y)$. Then either $x_* \notin \mathbb{X}(T, y, 0, 0)$ or $x_* \notin \mathbf{X}(T, 0)X_0$. The first is impossible, since $\mathbb{X}(T, y, 0, 0)$ is among $\mathbb{X}(T, y, P_0, \alpha)$ when the parameters $P_0 = 0$ and $\alpha = 0$ are chosen. Consider the second possibility. If $x_* \notin \mathbf{X}(T, 0)X_0$ is true, then $x_0 = \mathbf{X}(0, T)x_* \notin X_0$. By Lemma 2 and Remark 1, there exists an ellipsoid $E(Q, 0)$ containing X_0 but not containing x_0. There is also a parameter α such that $x_* \notin \mathbb{X}(T, y, Q, \alpha)$.

We get a contradiction, since the set includes only vectors x_*, for which $x_0'Qx_0 \leqslant 1$ and $x_* \in \mathbb{X}(T, y, 0, 0)$. The embedding

$$\bigcap_{P_0 \in \mathcal{P}_0, \alpha} \mathbb{X}(T, y, P_0, \alpha) \supset \mathcal{X}(T, y)$$

is obvious. □

4. Numerical examples

4.1. Double integrator

Consider the one-dimensional equations of motion of a material point

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = 0.$$

The set of possible initial states is a square:

$$X_0 = \{x_0 : |x_0^1| \leqslant 1, |x_0^2| \leqslant 1\}.$$

The measurement $y(t)$ are related to the state vector via the observation equation

$$y = x^1 + w(t),$$

where $w(t)$ is the measurement noise satisfying the integral constraint

$$\int_0^T w^2(t)dt \leqslant 1.$$

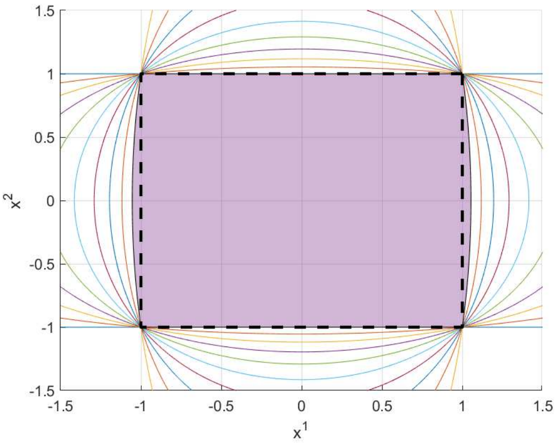


Figure 1. The set of possible initial states (black dashed line) and its approximation by external ellipses (pink fill).

Fig. 1 shows the approximation of the set of possible initial states by the intersection of a one-parameter family of ellipsoids with diagonal matrices $P_0 = [a, 0; 0, 1-a]$, where $a \in (0, 1)$. The intersection of the family of ellipses does not perfectly approximate the set of initial states, which is a square; to avoid this, one should use degenerate ellipsoids. Then, each ellipse $\{x \mid x' P_0 x \leqslant 1\}$ will contain the square of initial states X_0, and their intersection will give an external approximation.

The parameters here are $G = [1, 0]$, $A = [0, 1; 0, 0]$, and $T = 2$. For illustration, let us choose the signal generated by the admissible function $w(t) = 0.8\cos(t)$ and the admissible initial state $x_0 = [1; -1]/2$. Fig. 2 shows an approximation of the IS by a set of ellipsoids. The approximation of the IS (the white area on the left side of Fig. 2) coincides with the exact IS (the pink area on the right side of Fig. 2). The exact IS is obtained by the intersection of the reachable set at the terminal time (the black dashed line) and the IS without constraints on the initial state (the red dashed line).

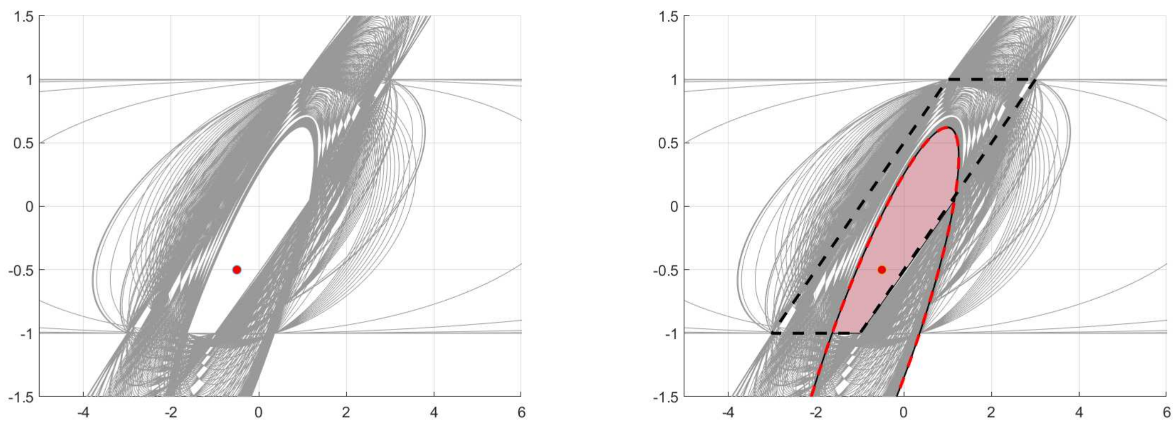


Figure 2. Double integrator. Approximation of the IS (on the left side) and the exact IS (on the right side). The red dot is the true state, the black dashed line is the reachable set at the terminal time ($T = 2$), and the red dashed line is the IS without constrains on the initial state.

4.2. Mathematical pendulum

Consider the equation

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = -25x^1.$$

The set of possible initial states is a circle: $X_0 = \{x_0 : |x_0| \leqslant 1\}$. The measurement equation is given by

$$y = x^1 + 0.8\cos(t),$$

where $w(t)$ is the measurement noise, for which

$$\int_0^T |y(t) - x^1(t)|^2 dt \leqslant 1$$

holds. The parameters here are $G = [1, 0]$, $A = [0, 1; -25, 0]$, and $T = 2$. The implementation of disturbances and the initial state coincide with those in the previous example: $w(t) = 0.8\cos(t)$ and $x_0 = [1; 1]/2$.

4.3. Double integrator II

Consider a one-dimensional motion of a material point under disturbances $w^1(t)$ [2]:

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = w^1(t), \quad 0 \leqslant t \leqslant T.$$

Let the disturbances w^1 also affect the measurement equation $y(t) = x^1(t) + w^1(t) + w^2(t)$, where w^2 is the measurement noise. Unfortunately, the calculation in [2] is inaccurate. Therefore, we need perform a new one. Define $w^1 - w^2$ by $z(t)$. Since $w^1 + w^2 = y - x^1$, we obtain the following equations:

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = \left(y - x^1 + z(t)\right)/2. \tag{4.1}$$

The vector-valued function $w(t)$ subjects to the integral constraint (2.2) with

$$R = I_2, \quad \mathbf{V} = \mathbb{R}^2, \quad X_0 = \left\{x \in \mathbb{R}^2 \mid |x_0^1| \leqslant 1,\, |x_0^2| \leqslant 1\right\},$$

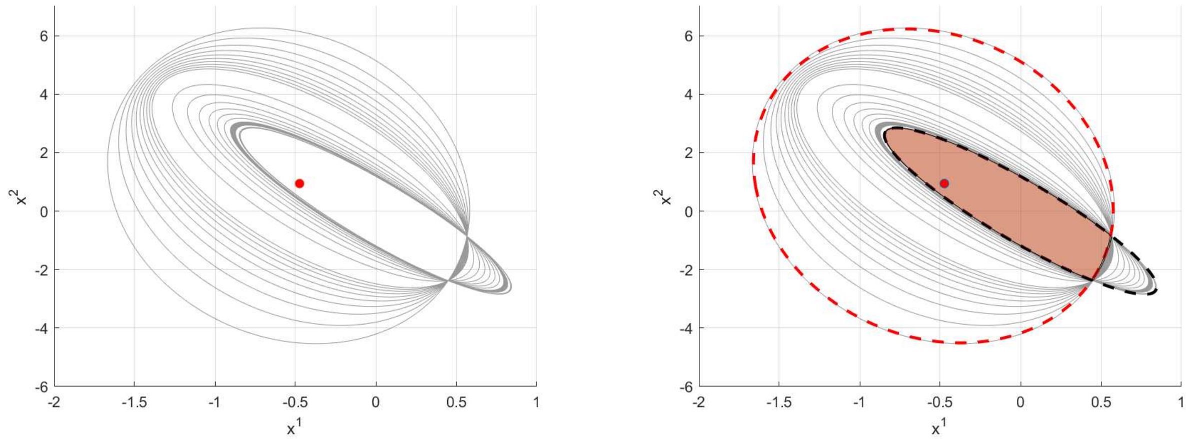


Figure 3. Mathematical pendulum. Approximation of the IS (left) and the exact IS (right). The red dot is the true state, the black dashed line is the reachable set at the terminal time ($T = 2$), and the red dashed line is the IS without constrains on the initial state.

i.e., this is the case of the absence of geometric constraints on $w(t)$. Since

$$(w^1 + w^2)^2 + z^2 = 2|w|^2,$$

inequality (2.2) takes the form

$$J(T, x_T, v, y) = \int_0^T \left(\left| y(t) - x^1(t) \right|^2 + z^2(t) \right) dt/2 \leqslant 1. \tag{4.2}$$

The constraints on initial states are the same as in the first example: a square is approximated by a one-parameter family of ellipses with diagonal matrices $P_0 = [a, 0; 0, 1 - a]$, where $a \in (0, 1)$. Then, each ellipse $\{x \mid x' P_0 x \leqslant 1\}$ contains the square of initial states X_0. Let is choose one more parameter $\alpha \in (0, 1)$ and consider the constraints

$$(1 - \alpha)|x_0|^2_{P_0} + \alpha J(T, x_T, v, y) < 1, \tag{4.3}$$

where J is defined in (4.2). The IS $\mathbf{X}_T^{a,\alpha}(y)$ for (4.1) under constrains (4.3) will contain the original IS $\mathbf{X}_T(y)$ for any signal in the original system. We will use relations (3.2). Then, we have

$$\mathbf{X}_T^{a,\alpha}(y) = \left\{ x_T \mid |x_T - \hat{x}(T)|^2_{P(T)} + h(T) < 1 \right\},$$

$$\dot{P} = -P(t)\tilde{A} - \tilde{A}' P(t) + \alpha G' G/2 - P(t) b C_1 b' P(t)/\alpha, \quad P(0) = (1 - \alpha) P_0;$$

$$\dot{\hat{x}}(t) = A\hat{x}(t) + \alpha \left(bc' + P^{-1}(t) G' \right) (y(t) - \hat{x}^1(t))/2,$$

$$\dot{h}(t) = \alpha \left| y(t) - \hat{x}^1(t) \right|^2 /2.$$

The parameters here are the same as in [2]: $b = [0, 0; 1, 0]$, $c = [1, 1]$, $G = [1, 0]$, $C = 1/2$, $C_1 = [1, -1; -1, 1]/2$, and $\tilde{A} = [0, 1; -0.5, 0]$. We take the signal generated by the admissible functions $w^1(t) = 0.8\cos(t)$ and $w^2(t) = 0.8\sin(t)$ and the admissible initial state $x_0 = [1; -1]/2$.

5. Conclusion

The problem of estimating the state vector for a linear autonomous system under uncertainty has been solved. For such systems, the IS can be obtained as an intersection of ellipsoids. The third example shows that this can be also true for systems with disturbances in the dynamics equation. The issue will be considered in subsequent works.

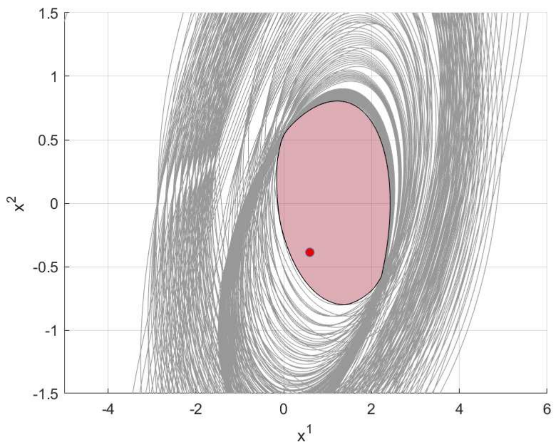


Figure 4. Double integrator II. The red dot is the true state at time $T = 3$.

REFERENCES

1. Ananyev B. I., Yurovskih P. A. On the approximation of estimation problems for controlled systems. *AIP Conf. Proc.*, 2019. Vol. 2164, No. 1. Art. no. 110001. P. 1–9. DOI: 10.1063/1.5130846
2. Ananyev B. I., Yurovskih P. A. Approximation of a guaranteed estimation problem with mixed constraints. *Trudy Inst. Mat. Mekh. UrO RAN*, 2020. T. 26, No. 4. P. 48–63. (in Russian) DOI: 10.21538/0134-4889-2020-26-4-48-63
3. Bertsekas D. P., Rhodes I. B. Recursive state estimation for a set-membership description of uncertainty. *IEEE Trans. Inform. Theory*, 1971. Vol. 16, No. 2. P. 117–128. DOI: 10.1109/TAC.1971.1099674
4. Chernousko F. L. *State Estimation for Dynamic Systems*. CRC Press: Boca Raton, 1994. 320 p.
5. Dontchev A. L., Rockafellar R. T. *Implicit Functions and Solution Mappings*. 2nd ed. Springer Ser. Oper. Res. Financ. Eng. New York: Springer, 2014. 466 p. DOI: 10.1007/978-1-4939-1037-3
6. Filippova T. F. External estimates for reachable sets of a control system with uncertainty and combined nonlinearity. *Proc. Steklov Inst. Math.*, 2018. Vol. 301, Suppl. 1. P. 32–43. DOI: 10.1134/S0081543818050036
7. Krasovskii N. N. *Teoriya upravleniya dvizheniem* [Theory of Control of Motion]. Moscow: Nauka, 1968. 476 p. (in Russian)
8. Kurzhanski A. B. *Upravlenie i nablyudenie v usloviyah neopredelennosti* [Control and Observation under Conditions of Uncertainty]. Moscow: Nauka, 1977. 392 p. (in Russian)
9. Kurzhanski A. B., Varaiya P. *Dynamics and Control of Trajectory Tubes: Theory and Computation.* Systems Control Found. Appl., vol. 85. Basel: Birkhäuser, 2014. 445 p. DOI: 10.1007/978-3-319-10277-1
10. Schweppe F. C. *Uncertain Dynamic Systems.* New Jersey: Prentice-Hall, Englewood Cliffs, 1973. 563 p.

URAL MATHEMATICAL JOURNAL, Vol. 7, No. 1, 2021, pp. 168–177
DOI: 10.15826/umj.2021.1.013

THE ASYMPTOTICS OF A SOLUTION OF THE MULTIDIMENSIONAL HEAT EQUATION WITH UNBOUNDED INITIAL DATA[1]

Sergey V. Zakharov

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences,
16 S. Kovalevskaya str., Ekaterinburg, 620108 Russia

svz@imm.uran.ru

Abstract: For the multidimensional heat equation, the long-time asymptotic approximation of the solution of the Cauchy problem is obtained in the case when the initial function grows at infinity and contains logarithms in its asymptotics. In addition to natural applications to processes of heat conduction and diffusion, the investigation of the asymptotic behavior of the solution of the problem under consideration is of interest for the asymptotic analysis of equations of parabolic type. The auxiliary parameter method plays a decisive role in the investigation.

Key words: Multidimensional heat equation, Cauchy problem, Asymptotics, Auxiliary parameter method.

1. Introduction

In 1822, J. Fourier published his most fundamental work [4], where the heat conduction equation was presented and analyzed. This event provided a strong impetus for later researches in the fields of partial differential equations and trigonometric series. The famous equation has been further successfully used for effective descriptions of molecular diffusion, stochastic motion, the capillary conduction of liquids in porous media, and even for the analysis of social economic data. Already Fourier himself pointed out the universality of this mathematical model *sine qua non* in his eminent book as follows: “Il est facile de juger combien ces recherches intéressent les sciences physiques et l'économie civile, et quelle peut être leur influence sur les progrès des arts qui exigent l'emploi et la distribution du feu.”[2] Fourier's preliminary theoretical studying of heat phenomena and some vivid particulars of his elaborations in early 1800s are expressively reflected in the prefatory part of [4]. The historical survey [10] supplied with appropriate general and specialized references depicts many significant details of the subsequent life of the heat equation during the XIX and XX centuries.

Since the literature about the heat equation, in particular, and parabolic equations, in general, is immense, it is impossible in this introduction to give a complete picture of available results, and the bibliography below is of course by no means exhaustive. Here, we mention that existence and uniqueness theorems were obtained for a wide class of parabolic equations and systems [6, 15, 18, 19]; some results for unbounded solutions were presented in [11, 13]. As for the long-time behavior of solutions, we see that their stabilization, certain estimates, and the leading terms of asymptotics

[1]Dedicated to the 200th anniversary of Charles Hermite and “Théorie analytique de la chaleur” by Joseph Fourier.

[2]“It is easy to judge how much these researches are interesting for the physical sciences and the civil economy and what may be their influence on the progress of the arts which require the employment and the distribution of fire.”

were mainly considered [2, 8, 12, 17]. Complete asymptotic expansions of solutions into infinite series in inverse integer powers of the time variable were earlier obtained by Friedman in [5] and [6, Ch. 6] for bounded space-domains.

In the present paper, the long-time asymptotics of the solution of the Cauchy problem for the multidimensional heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x_1^2} + \ldots + \frac{\partial^2 u}{\partial x_m^2}, \quad t > 0, \quad m \geqslant 2, \tag{1.1}$$

$$u(x_1, \ldots, x_m, 0) = \Lambda(x_1, \ldots, x_m), \quad (x_1, \ldots, x_m) \in \mathbb{R}^m, \tag{1.2}$$

is obtained for a locally Lebesgue integrable initial function $\Lambda : \mathbb{R}^m \to \mathbb{R}$ of polynomial growth. As is well known [18], in the class of smooth functions of moderate growth for $t > 0$, there exists a unique solution of problem (1.1)–(1.2) and it can be written in the form of the Poisson integral[3]

$$u(x, t) = \frac{1}{(4\pi t)^{m/2}} \int\limits_{\mathbb{R}^m} \Lambda(s) \exp\Big(- \frac{|s - x|^2}{4t} \Big) ds, \tag{1.3}$$

where $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$, $s = (s_1, \ldots, s_m) \in \mathbb{R}^m$, and $ds = ds_1 \ldots ds_m$.

It should be noted that the investigation of the asymptotic behavior of the function $u(x, t)$, in addition to possible natural applications to the modeling of physical processes of heat conduction and diffusion, may be of interest for the asymptotic analysis of solutions of nonlinear parabolic equations by the matching method [9, 21] as well as for the theory of invariants [7] and some issues of matrix geometry [14].

Below, a complete asymptotic expansion of the solution $u(x, t)$ of problem (1.1)–(1.2) is found as $|x| + t \to +\infty$ under the following suppositions:

$$\Lambda(x_1, \ldots, x_m) = 0, \quad x_1 < 0, \tag{1.4}$$

$$\Lambda(x_1, \ldots, x_m) = x_1^p \sum_{n=0}^{\infty} x_1^{-n} \sum_{j=0}^{n} \Lambda_{n,j}(x') \ln^j x_1, \quad x_1 \to +\infty, \tag{1.5}$$

where p is a positive integer and $\Lambda_{n,j}(x')$ are Lebesgue integrable functions of $x' = (x_2, \ldots, x_m)$; for simplicity, we also suppose that

$$\begin{aligned} \operatorname{supp} \Lambda &\subset \big\{ (x_1, \ldots, x_m) : x_1 > 0,\ |x_2| + \ldots + |x_m| < x_1^{\nu} \big\}, \quad \nu > 0, \\ \operatorname{supp} \Lambda_{n,j} &\subset \big\{ (x_2, \ldots, x_m) : |x_2| + \ldots + |x_m| < r_n \big\}, \quad r_n > 0. \end{aligned} \tag{1.6}$$

Although Λ is a function of several variables, the asymptotic series (1.5) must be understood here in the usual sense of Poincaré [16, § 1] due to the second condition (1.6), that is

$$\Lambda(x_1, \ldots, x_m) = \sum_{n=0}^{N-1} x_1^{p-n} \sum_{j=0}^{n} \Lambda_{n,j}(x') \ln^j x_1 + O\big(x_1^{p-N} \ln^N x_1\big), \quad x_1 \to +\infty, \tag{1.7}$$

for any integer $N \geqslant 1$. It should be also said that the appearance of asymptotic series of form (1.5) is typical for the matching method [9].

The main difficulty of the calculation of the asymptotic expansion of integral (1.3) is exactly due to condition (1.5) and the "smearing" of the integrand exponent as $t \to +\infty$; if we formally put $t = +\infty$, then we generally get the divergence of the integral. Thus, the asymptotic limit under consideration is diametrically opposite to the well-known case of the integrals of Laplace's type with the sharpening exponent and a suitable computational technique suggested by Danilin in [1] is therefore complementary to the standard Laplace method. This technique is called the auxiliary parameter method.

[3]In essence, this solution was given by Fourier [4, Ch. IX, §392].

2. Applying the auxiliary parameter method

To obtain the asymptotic behavior of integral (1.3) as the space-time variables (x,t) independently tend to infinity, we apply a scheme similar to that used in [20] for the solution of the heat equation in $\mathbb{R}^1_x \times \mathbb{R}^+_t$. First of all, taking into account condition (1.4), we represent function (1.3) in the form of the sum

$$u(x,t) = U_0(x,t) + U_1(x,t), \tag{2.1}$$

where

$$U_0(x,t) = \int\limits_0^{\sigma(x,t)} \int\limits_{\mathbb{R}^{m-1}} \ldots \, ds' ds_1, \quad U_1(x,t) = \int\limits_{\sigma(x,t)}^{+\infty} \int\limits_{\mathbb{R}^{m-1}} \ldots \, ds' ds_1,$$

$$\sigma(x,t) = \left(|x|^2 + t\right)^{\beta/2}, \quad 0 < \beta < 1, \tag{2.2}$$

the dots denote the integrand in formula (1.3) together with the factor $(4\pi t)^{-m/2}$, the number β is an arbitrary parameter, and $ds' = ds_2 \ldots ds_m$. Under conditions (1.4) and (1.5), the asymptotics of the integrals $U_0(x,t)$ and $U_1(x,t)$ can be computed by using the expansions of the kernel exponent and the initial function Λ, respectively.

2.1. Asymptotics of $U_1(x,t)$

In the integral $U_1(x,t)$, we make the change $s_1 = 2z\sqrt{t}$ and put

$$\mu(x,t) = \frac{\sigma(x,t)}{2\sqrt{t}}, \quad \eta_1 = \frac{x_1}{2\sqrt{t}}. \tag{2.3}$$

Next, using condition (1.5), for any integer $N \geqslant p+1$, we obtain (hereinafter we often omit the arguments of σ and μ)

$$\begin{aligned} U_1(x,t) &= \frac{1}{\pi^{m/2}(4t)^{(m-1)/2}} \int\limits_\mu^{+\infty} \exp\left(-(\eta_1 - z)^2\right) \int\limits_{\mathbb{R}^{m-1}} \Lambda(2z\sqrt{t}, s') \exp\left(-\frac{|s'-x'|^2}{4t}\right) ds' dz \\ &= \frac{t^{p/2}}{\sqrt{\pi}} \sum_{n=0}^{N-1} 2^{p-n} t^{-n/2} \sum_{j=0}^{n} \int\limits_\mu^{+\infty} z^{p-n} \ln^j(2z\sqrt{t}) \, \exp\left(-(z-\eta_1)^2\right) dz \\ &\times \frac{1}{(4\pi t)^{(m-1)/2}} \int\limits_{\mathbb{R}^{m-1}} \Lambda_{n,j}(s') \exp\left(-\frac{|s'-x'|^2}{4t}\right) ds' + R(x,t), \end{aligned}$$

where

$$|R(x,t)| \leqslant \frac{M_N}{\sqrt{t}} \int\limits_\sigma^{+\infty} s_1^{p-N} \ln^N s_1 \exp\left(-\frac{(s_1-x_1)^2}{4t}\right) ds_1, \quad M_N > 0,$$

by formula (1.7). Then, for $N \geqslant p+1$, we have

$$\begin{aligned} U_1(x,t) &= \frac{t^{p/2}}{\sqrt{\pi}} \sum_{n=0}^{N-1} 2^{p-n} t^{-n/2} \sum_{j=0}^{n} \sum_{l=0}^{j} \frac{j! \ln^l t}{2^l l!(j-l)!} \int\limits_\mu^{+\infty} z^{p-n} \ln^{j-l}(2z) \, \exp\left(-(z-\eta_1)^2\right) dz \\ &\times \frac{1}{(4\pi t)^{(m-1)/2}} \int\limits_{\mathbb{R}^{m-1}} \Lambda_{n,j}(s') \exp\left(-\frac{|s'-x'|^2}{4t}\right) ds' + O\left(\sigma^{p-N} \ln^N \sigma\right) \end{aligned}$$

as $\sigma = \sigma(x,t) \to +\infty$. Changing the order of summation, we find

$$\begin{aligned} U_1(x,t) = \, & t^{p/2} \sum_{n=0}^{N-1} t^{-n/2} \sum_{l=0}^{n} \ln^l t \sum_{j=l}^{n} \frac{j!\, 2^{p-n-l}}{\sqrt{\pi}\, l!(j-l)!} \int\limits_{\mu}^{+\infty} z^{p-n} \ln^{j-l}(2z) \exp\left(-(z-\eta_1)^2\right) dz \\ & \times \frac{1}{(4\pi t)^{(m-1)/2}} \int\limits_{\mathbb{R}^{m-1}} \Lambda_{n,j}(s') \exp\left(-\frac{|s'-x'|^2}{4t}\right) ds' + O\left(\sigma^{p-N} \ln^N \sigma\right), \quad \sigma \to +\infty. \end{aligned} \tag{2.4}$$

To handle the integral with respect to z, it is convenient to consider first the following set of independent variables:

$$T_\alpha = \big\{(x,t) \,:\, x \in \mathbb{R}^m,\ t \geqslant |x|^\alpha > 1,\ 1+\beta < \alpha < 2\big\}. \tag{2.5}$$

The obvious inequalities

$$\sigma(x,t) \leqslant (t^{2/\alpha} + t)^{\beta/2} < 2^{\beta/2} t^{\beta/\alpha}$$

for $(x,t) \in T_\alpha$ imply that

$$t > 2^{-\alpha/2} [\sigma(x,t)]^{\alpha/\beta} \quad \text{for } (x,t) \in T_\alpha; \tag{2.6}$$

therefore, on account of the first definition (2.3), we obtain

$$0 < \mu(x,t) < 2^{\alpha/4-1} [\sigma(x,t)]^{-\gamma} \quad \text{for} \quad (x,t) \in T_\alpha, \quad \text{where} \quad \gamma = \frac{\alpha}{2\beta} - 1 > 0. \tag{2.7}$$

For $0 \leqslant n \leqslant p$, we have

$$\begin{aligned} \int\limits_{\mu}^{+\infty} z^{p-n} \ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz &= \int\limits_{0}^{+\infty} z^{p-n} \ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz - \int\limits_{0}^{\mu} z^{p-n} \ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz \\ &= \int\limits_{-\eta_1}^{+\infty} (\eta_1 + s)^{p-n} \ln^{j-l}(2(\eta_1 + s))\, e^{-s^2} ds - \int\limits_{0}^{\mu} z^{p-n} \ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz. \end{aligned}$$

Since by (2.7) $\mu \to +0$ as $\sigma \to +\infty$ for $(x,t) \in T_\alpha$, it follows that

$$\begin{aligned} \int\limits_{\mu}^{+\infty} z^{p-n} \ln^{j-l}(2z) e^{-(z-\eta_1)^2} dz = \, & \sum_{q=0}^{p-n} \frac{(p-n)!\, \eta_1^{p-n-q}}{q!(p-n-q)!} \int\limits_{-\eta_1}^{+\infty} s^q \ln^{j-l}[2(\eta_1+s)] e^{-s^2}\, ds \\ & + e^{-\eta_1^2} \sum_{s:\, r_s^2 + l_s^2 \neq 0} b'_s \eta_1^{n_s} \mu^{r_s} \ln^{l_s} \mu + O\left(\sigma^{-\gamma N}\right), \quad \sigma \to +\infty, \end{aligned} \tag{2.8}$$

where the finite sum over s with b'_s being some constants and n_s, r_s, l_s being some nonnegative integers depends naturally on N. For $n > p$, we have

$$\begin{aligned} \int\limits_{\mu}^{+\infty} z^{p-n} \ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz &= \int\limits_{1}^{+\infty} z^{p-n} \ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz \\ &+ \int\limits_{\mu}^{1} \ln^{j-l}(2z) \Psi_{n-p}(z,\eta_1) dz + e^{-\eta_1^2} \sum_{r=0}^{n-p-1} P_r(\eta_1) \int\limits_{\mu}^{1} \ln^{j-l}(2z) z^{r+p-n} dz, \end{aligned}$$

where $P_r(\eta_1)$ are some polynomials of degree r,

$$\Psi_{n-p}(z,\eta_1) = z^{p-n}\Big[e^{-(z-\eta_1)^2} - e^{-\eta_1^2}\sum_{r=0}^{n-p-1} H_r(\eta_1)\frac{z^r}{r!}\Big], \tag{2.9}$$

and the sum in the square brackets is a partial sum of the Maclaurin series for the function $\exp(2z\eta_1 - z^2)$ in variable z with $H_r(\eta_1)$ being the Hermite polynomials of degree r. This implies the equality

$$\begin{aligned} &\int_{\mu}^{+\infty} z^{p-n}\ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz \\ &= \int_{1}^{+\infty} z^{p-n}\ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz + e^{-\eta_1^2}\sum_{s:\, r_s^2+l_s^2\neq 0} b_s''\, \eta_1^{n_s}\mu^{r_s}\ln^{l_s}\mu \\ &+ \int_0^1 \ln^{j-l}(2z)\Psi_{n-p}(z,\eta_1)dz - \int_0^{\mu}\ln^{j-l}(2z)\Psi_{n-p}(z,\eta_1)dz \end{aligned} \tag{2.10}$$

with b_s'' being some constants and n_s, r_s, l_s being some nonnegative integers. From formula (2.9) we easily conclude that the function $\Psi_{n-p}(z,\eta_1)$ has no singularities as $z\to 0$; therefore, the last two integrals in (2.10) converge and relation (2.10) itself thus becomes

$$\int_{\mu}^{+\infty} z^{p-n}\ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz = J_{p,n,j,l}(\eta_1) + e^{-\eta_1^2}\sum_{s:\, r_s^2+l_s^2\neq 0} b_s'''\, \eta_1^{n_s}\mu^{r_s}\ln^{l_s}\mu + O\left(\sigma^{-\gamma N}\right) \tag{2.11}$$

as $\sigma\to+\infty$, where

$$J_{p,n,j,l}(\eta_1) = \int_1^{+\infty} z^{p-n}\ln^{j-l}(2z)\, e^{-(z-\eta_1)^2} dz + \int_0^1 \ln^{j-l}(2z)\Psi_{n-p}(z,\eta_1)dz, \tag{2.12}$$

b_s''' are some constants, n_s, r_s, l_s are some nonnegative integers, and γ is defined in (2.7).

Using the second condition (1.6) and Maclaurin's expansion for the exponent in the integrand of (2.4) in $s't^{-1/2}$, for any natural $N^*\geqslant 1$, we obtain

$$\begin{aligned} &\frac{1}{(4\pi t)^{(m-1)/2}}\int_{\mathbb{R}^{m-1}} \Lambda_{n,j}(s')\exp\Big(-\frac{|s'-x'|^2}{4t}\Big)ds' \\ &= t^{(1-m)/2}\exp\left(-|\eta'|^2\right)\Big[\sum_{l=0}^{N^*-1} t^{-l/2}Q_l^{(n,j)}(\eta') + O\Big(t^{-N^*/2}|\eta'|^{N^*}\Big)\Big], \end{aligned} \tag{2.13}$$

where $Q_l^{(n,j)}(\eta')$ are some lth degree polynomials in $\eta' = 2^{-1}t^{-1/2}x'$ whose coefficients depend on n and j. Substituting expressions (2.8), (2.11), and (2.13) into formula (2.4) and taking into account that $\sigma^{-(\gamma+\alpha/2\beta)N} = O(\sigma^{-N})$, since $\gamma+\alpha/2\beta = \alpha/\beta - 1 > 1$, we find that

$$U_1(x,t) = t^{(p-m+1)/2}\sum_{n=0}^{N-1} t^{-n/2}\sum_{l=0}^{n}\widetilde{S}_{n,l}(\eta)\ln^l t + V_{1,N}(\mu,\eta,t) + O\left(\sigma^{p-N}\ln^N\sigma\right) \tag{2.14}$$

as $\sigma \to +\infty$, where, according to formulas (2.9) and (2.12), the coefficients $\widetilde{S}_{n,l}(\eta)$ are some smooth functions of polynomial growth for $0 \leqslant n \leqslant p$ and of superexponential decreasing for $n > p$,

$$V_{1,N}(\mu,\eta,t) = \exp\left(-|\eta|^2\right) \sum_{s:\, r_s^2+l_s^2\neq 0} a'_s\, t^{k_s} \eta^{n_s} \mu^{r_s} \ln^{l_s} \mu \tag{2.15}$$

is a finite sum with $\eta^{n_s} = \eta_1^{n_{1,s}} \dots \eta_m^{n_{m,s}}$, a'_s being some real constants, k_s being half-integer numbers, and $n_{j,s}$, r_s, l_s being some nonnegative integers. Because of the factor $\exp(-|\eta|^2)$, the estimate of the remainder in formula (2.14) remains true for the values of the independent variables from the set

$$X_\alpha = \big\{(x,t) \,:\, |x| > 1,\ 0 < t < |x|^\alpha\big\}, \tag{2.16}$$

since for $(x,t) \in X_\alpha$ there hold the following inequalities:

$$\mu^2 = \frac{(|x|^2+t)^\beta}{4t} < 2|\eta|^2|x|^{-2(1-\beta)}, \quad |\eta|^2 > \frac{1}{4}|x|^{2-\alpha} > \frac{1}{8}\sigma^{(2-\alpha)/\beta}. \tag{2.17}$$

2.2. Asymptotics of $U_0(x,t)$

Now, let us pass to the evaluation of the integral

$$U_0(x,t) = \frac{1}{(4\pi t)^{m/2}} \int\limits_0^{\sigma(x,t)} ds_1 \int\limits_{\mathbb{R}^{m-1}} ds'\, \Lambda(s_1,s') \exp\left(-\frac{|s-x|^2}{4t}\right) ds.$$

From the obvious inequality $|x|^2 \leqslant [\sigma(x,t)]^{2/\beta}$ and inequality (2.6) we conclude that

$$\frac{|s|^2}{t} = O(\sigma^{-2\delta}), \quad \frac{x_k s_k}{t} = \frac{2\eta_k s_k}{\sqrt{t}} = O(\sigma^{-\delta}), \quad \delta = \frac{\alpha-1}{\beta} - 1 > 0, \tag{2.18}$$

for $|s| \leqslant \sigma$ and $(x,t) \in T_\alpha$, where $1 \leqslant k \leqslant m$. Then, using conditions (1.6), (1.7) and estimates (2.18), we represent the integral $U_0(x,t)$ in the following form:

$$U_0(x,t) = \frac{\exp(-|\eta|^2)}{(4\pi t)^{m/2}} \Bigg[\int\limits_0^{\sigma} \int\limits_{\mathbb{R}^{m-1}} \Lambda(s_1,s') \sum_{q=0}^{N-1} \frac{1}{q!} \left(\frac{\eta_1 s_1 + \ldots + \eta_m s_m}{\sqrt{t}} - \frac{|s|^2}{4t} \right)^q ds' ds_1 + O\Big(\sigma^{p+1-\delta N}\Big) \Bigg]$$

as $\sigma \to +\infty$ with any $N \geqslant 1$. Because of the factor $\exp(-|\eta|^2)$, the estimate of the remainder holds also true on the set X_α defined by (2.16). Expanding the parenthesis in the above formula for $U_0(x,t)$ and changing the order of summation, we obtain

$$\begin{gathered} U_0(x,t) = \frac{\exp(-|\eta|^2)}{t^{m/2}} \sum_{n=0}^{N-1} t^{-n/2} \sum_{\substack{0\leqslant k_1+\ldots+k_m\leqslant n \\ 0\leqslant l_1+l_{2,2}+\ldots+l_{2,m}\leqslant n}} a_{k,l}\, \eta^k \int\limits_0^{\sigma} \int\limits_{\mathbb{R}^{m-1}} s_1^{l_1} (s')^{l_2} \Lambda(s_1,s') ds' ds_1 \\ + O\Big(\frac{\exp(-|\eta|^2)}{t^{m/2}} \sigma^{p+1-\delta N}\Big) \end{gathered}$$

as $\sigma \to +\infty$, where $a_{k,l} = a_{k_1,\ldots,k_m,l_1,l_{2,2},\ldots,l_{2,m}}$ are some constants, $\eta^k = \eta_1^{k_1} \dots \eta_m^{k_m}$, and $(s')^{l_2} = s_2^{l_{2,2}} \dots s_m^{l_{2,m}}$. Keeping in mind the asymptotic condition (1.7), we transform the multiple integral appeared above as follows:

$$\int\limits_0^{\sigma} \int\limits_{\mathbb{R}^{m-1}} s_1^{l_1} (s')^{l_2} \Lambda(s_1,s')\, ds' ds_1 = \int\limits_0^{1} \int\limits_{\mathbb{R}^{m-1}} s_1^{l_1} (s')^{l_2} \Lambda(s_1,s')\, ds' ds_1$$

$$
\begin{aligned}
&+\int_1^\sigma \int_{\mathbb{R}^{m-1}} s_1^{l_1}(s')^{l_2}\Big[\Lambda(s_1,s') - \sum_{q=0}^{p+l_1+1} s_1^{p-q}\sum_{j=0}^{q}\Lambda_{q,j}(s')\ln^j s_1\Big]ds'ds_1\\
&\qquad+\int_1^\sigma \int_{\mathbb{R}^{m-1}} \Big[(s')^{l_2}\sum_{q=0}^{p+l_1+1} s_1^{p-q+l_1}\sum_{j=0}^{q}\Lambda_{q,j}(s')\ln^j s_1\Big]ds'ds_1\\
&= \sum_{j=0}^{p+l_1+1} c_{l_1,l_2,j}\ln^{j+1}\sigma + \sum_{i,j:\, i\neq 0} c^*_{l_1,l_2,i,j}\,\sigma^i\ln^j\sigma + O\big(\sigma^{-N^*}\ln^{N^*}\sigma\big), \quad \sigma\to+\infty,
\end{aligned}
$$

with $c_{l_1,l_2,j}$ and $c^*_{l_1,l_2,i,j}$ being some constants, where the finite sum over i,j depends naturally on a sufficiently large N^*; here we used the elementary relation

$$
\int_1^\sigma s_1^k\ln^j s_1 ds_1 = \sigma^{k+1}\sum_{l=0}^{j-1}\frac{(-1)^l j!\ln^{j-l}\sigma}{(k+1)^{l+1}(j-l)!} + (\sigma^{k+1}-1)\frac{(-1)^j j!}{(k+1)^{j+1}} \quad (k\geqslant 0,\quad j\geqslant 1).
$$

From formulas (2.3), inequality (2.6), the uniform estimate

$$
t^{-m/2}\eta^n\exp\big(-|\eta|^2\big) = O\big(\sigma^{-\alpha m/2\beta}\big),
$$

and the previous asymptotic expression for $U_0(x,t)$, it follows that

$$
U_0(x,t) = \frac{\exp\big(-|\eta|^2\big)}{t^{m/2}}\sum_{n=0}^{N-1} t^{-n/2}\sum_{j=0}^{p+n+2}\Pi_{n,j}(\eta)\ln^j t + V_{0,N}(\mu,\eta,t) + O\big(\sigma^{p+1-\delta N}\big) \tag{2.19}
$$

as $\sigma\to+\infty$, where δ is defined in (2.18), $\Pi_{n,j}(\eta)$ are some polynomials of degree n, and the finite sum

$$
V_{0,N}(\mu,\eta,t) = \exp\big(-|\eta|^2\big)\sum_{s:\, r_s^2+l_s^2\neq 0} a''_s t^{k_s}\eta^{n_s}\mu^{r_s}\ln^{l_s}\mu, \tag{2.20}
$$

with a''_s being some constants, is obtained similarly to expression (2.15).

2.3. Evaluation of the "virtual terms"

In the sequel, it is convenient to suppose that $1+\beta<\alpha<1+2\beta$, whence we find the inequalities $0<\delta=(\alpha-1)/\beta-1<1$ and the asymptotic estimate $\sigma^{p-N}\ln^N\sigma = O\left(\sigma^{p+1-\delta N}\right)$ as $\sigma\to+\infty$. Then substituting expansions (2.14) and (2.19) into formula (2.1), we summarize the results of the previous two subsections as follows.

Lemma 1. *For the solution of the Cauchy problem (1.1)–(1.2), the asymptotic formula*

$$
\begin{aligned}
u(x,t) = t^{-m/2}\sum_{n=0}^{N-1} t^{-n/2}\Big[\sum_{l=0}^{n} t^{(p+1)/2}\widetilde{S}_{n,l}(\eta)\ln^l t + \sum_{j=0}^{p+n+2}\Pi_{n,j}(\eta)\exp\big(-|\eta|^2\big)\ln^j t\Big]\\
+V_{0,N}(\mu,\eta,t)+V_{1,N}(\mu,\eta,t)+O\big(\sigma^{p+1-\delta N}\big)
\end{aligned} \tag{2.21}
$$

holds true as $\sigma\to+\infty$, where $N\geqslant p+1$, $\widetilde{S}_{n,j}(\eta)$ are smooth functions of polynomial growth, $\Pi_{n,j}(\eta)$ are nth degree polynomials, and the functions $V_{0,N}(\mu,\eta,t)$ and $V_{1,N}(\mu,\eta,t)$ are defined by expressions (2.15) and (2.20).

Now we must evaluate the "virtual terms" that depend on the value $\mu(x,t) = 2^{-1}t^{-1/2}(|x|^2+t)^{\beta/2}$ with the arbitrary parameter β.

From inequalities (2.17), we conclude that, for $(x,t) \in X_\alpha$, any integer numbers $n_{s,j}$, r_s, l_s, and half-integer number k_s, there exist $C>0$ and $q>0$ such that

$$\left|t^{k_s}\eta_j^{n_{s,j}}\mu^{r_s}\ln^{l_s}\mu\right|\exp\left(-|\eta|^2\right) \leqslant C\sigma^q\exp\left(-8^{-1}\sigma^{(2-\alpha)/\beta}\right).$$

Consequently, the expressions $V_{0,N}(\mu,\eta,t)$ and $V_{1,N}(\mu,\eta,t)$ in formulas (2.14) and (2.19) are exponentially small for $(x,t)\in X_\alpha$, since $\alpha<2$ by (2.5).

For $(x,t)\in T_\alpha$, we introduce a small quantity $\varepsilon = (|x|^2+t)^{-1/4}$; whence, according to (2.2) and (2.3), we easily get the relations

$$\sigma = \varepsilon^{-2\beta}, \quad \mu = 2^{-1}t^{-1/2}\varepsilon^{-2\beta}. \tag{2.22}$$

Then, by formulas (2.15), (2.20), and (2.22), we have

$$V_{0,N}(\mu,\eta,t)+V_{1,N}(\mu,\eta,t) = \exp\left(-|\eta|^2\right)\sum_{s=1}^{L(N)} a_s'''\, t^{k_s}\ln^{k_s'}t\;\eta^{n_s}\varepsilon^{-2\beta r_s}\ln^{l_s}\varepsilon^{2\beta}, \tag{2.23}$$

where $\varepsilon\to+0$ as $|x|^2+t\to+\infty$, $L(N)\in\mathbb{N}$, a_s''' are some constants, $\eta^{n_s} = \eta_1^{n_{s,1}}\dots\eta_m^{n_{s,m}}$, k_s are half-integer numbers, k_s', $n_{s,j}$, r_s, l_s, are nonnegative integers such that $r_s^2+l_s^2\neq 0$, and β is an arbitrary parameter, without loss of generality, such that $0<\beta_1\leqslant\beta\leqslant\beta_2<1$, where $\beta_1<\beta_2$.

By virtue of the arbitrariness of the value β, from formulas (2.21) and (2.23) with $\beta=\beta_1$ and $\beta=\beta_2$ such that all numbers $2\beta_1 r_1,\dots,2\beta_1 r_{L(N)}$, $2\beta_2 r_1,\dots,2\beta_2 r_{L(N)}$ are pairwise distinct, we obtain the following asymptotic relation with $r_s^2+l_s^2\neq 0$:

$$\exp\left(-|\eta|^2\right)\sum_{s=1}^{L(N)} a_s'''\,t^{k_s}\ln^{k_s'}t\;\eta^{n_s}\left(\varepsilon^{-2\beta_1 r_s}\ln^{l_s}\varepsilon^{2\beta_1}-\varepsilon^{-2\beta_2 r_s}\ln^{l_s}\varepsilon^{2\beta_2}\right) = O\Big(\varepsilon^{2(\alpha-1-\beta_1)N-2\beta_1(p+1)}\Big)$$

as $\varepsilon\to+0$. Consequently, taking into account the finiteness of the sum in the left-hand side, we have to conclude about every particular term in the left-hand side that either its order is not greater than the estimate in the right-hand side or the corresponding coefficient a_s''' is equal to zero. Thus, we arrive at the following statement with $\beta=\beta_1$.

Lemma 2. *For some $\beta\in(0,1)$ and $\alpha\in(1+\beta,1+2\beta)$, the asymptotic estimate*

$$V_{0,N}(\mu,\eta,t)+V_{1,N}(\mu,\eta,t) = O\Big((|x|^2+t)^{-(\alpha-1-\beta)N/2+\beta(p+1)/2}\Big) \tag{2.24}$$

holds true as $|x|^2+t\to+\infty$.

3. Asymptotics of the solution

Immediately from Lemmas 1 and 2, we obtain our main result.

Theorem 1. *Let $u:\mathbb{R}^m\times\mathbb{R}^+\to\mathbb{R}$ be the solution of the Cauchy problem*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x_1^2}+\dots+\frac{\partial^2 u}{\partial x_m^2}, \quad t>0, \quad m\geqslant 2,$$

$$u(x_1,\dots,x_m,0) = \Lambda(x_1,\dots,x_m), \quad (x_1,\dots,x_m)\in\mathbb{R}^m,$$

with a locally Lebesgue integrable initial function $\Lambda : \mathbb{R}^m \to \mathbb{R}$. *And let the following conditions be fulfilled*:

$$\Lambda(x_1, \ldots, x_m) = 0 \quad \textit{for} \quad x_1 < 0,$$
$$\Lambda(x_1, \ldots, x_m) = x_1^p \sum_{n=0}^{\infty} x_1^{-n} \sum_{j=0}^{n} \Lambda_{n,j}(x_2, \ldots, x_m) \ln^j x_1 \quad \textit{as} \quad x_1 \to +\infty,$$

where p *is a positive integer,*

$$\operatorname{supp} \Lambda \subset \big\{(x_1, \ldots, x_m) : x_1 > 0,\ |x_2| + \ldots + |x_m| < x_1^{\nu}\big\}, \quad \nu > 0,$$
$$\operatorname{supp} \Lambda_{n,j} \subset \big\{(x_2, \ldots, x_m) : |x_2| + \ldots + |x_m| < r_n\big\}, \quad r_n > 0.$$

Then there holds the asymptotic formula

$$u(x_1, \ldots, x_m, t) = t^{-m/2} \sum_{n=0}^{\infty} t^{-n/2} \sum_{j=0}^{p+n+2} \ln^j t \left[t^{(p+1)/2} S_{n,j}(\eta_1, \ldots, \eta_m) + \Pi_{n,j}(\eta_1, \ldots, \eta_m) \exp\left(-|\eta|^2\right) \right]$$

as $|x_1| + \ldots + |x_m| + t \to +\infty$, *where* $S_{n,j}(\eta_1, \ldots, \eta_m)$ *are smooth functions of polynomial growth and* $\Pi_{n,j}(\eta_1, \ldots, \eta_m)$ *are* n*th degree polynomials in the self-similar variables*

$$\eta_1 = \frac{x_1}{2\sqrt{t}}, \quad \ldots, \quad \eta_m = \frac{x_m}{2\sqrt{t}}.$$

4. Conclusion

According to formulas (2.14), (2.19), and (2.24), the obtained expansion of the solution in Theorem 1 is understood in the sense of Erdélyi [3, Definition 2.4] with the gauge (asymptotic) sequence $\big\{(|x|^2 + t)^{-\rho N}\big\}_{N=1}^{\infty}$, where $\rho > 0$, that is

$$u(x,t) = \sum_{n=0}^{N-1} t^{-(m+n)/2} \sum_{j=0}^{p+n+2} \ln^j t \left[t^{(p+1)/2} S_{n,j}\Big(\frac{x}{2\sqrt{t}}\Big) + \Pi_{n,j}\Big(\frac{x}{2\sqrt{t}}\Big) \exp\Big(-\frac{|x|^2}{4t}\Big) \right]$$
$$+ O\left((|x|^2 + t)^{-\rho N}\right)$$

for each $N \geqslant p + 1$ as $|x|^2 + t \to +\infty$. In general, the exact formulas for $S_{n,j}(\eta)$ and $\Pi_{n,j}(\eta)$ are fairly cumbersome; however, by using the above proofs, one can derive them in particular cases. Note that, as shown by earlier investigations, asymptotic expansions in half-integer powers of t are naturally intrinsic to solutions of the heat equation, see, for example, [19, Ch. X, §1] and [20].

In conclusion, following Poincaré's thesis "sans généralisation, la prévision est impossible"[4] (see his "*La Science et l'Hypothèse*", Ch. IX), it is appropriate to say that the immense variety of asymptotics of initial data together with the account of possible external sources of heat opens a wide field of further investigation of the long-time behavior of heat distribution by the above-presented method; in addition, other types of equations whose solutions have the form of convolutions can also be treated in a similar way.

REFERENCES

1. Danilin A. R. Asymptotic behaviour of bounded controls for a singular elliptic problem in a domain with a small cavity. *Sb. Math.*, 1998. Vol. 189, No. 11. P. 1611–1642. DOI: 10.1070/SM1998v189n11ABEH000364

[4] "prevision is impossible without generalization"

2. Denisov V. N. On the behavior of solutions of parabolic equations for large values of time. *Russian Math. Surveys*, 2005. Vol. 60, No. 4. P. 721–790. DOI: 10.1070/RM2005v060n04ABEH003675
3. Erdélyi A., Wyman M. The asymptotic evaluation of certain integrals. *Arch. Rational Mech. Anal.*, 1963. Vol. 14, P. 217–260. DOI: 10.1007/BF00250704
4. Fourier J. *Théorie Analytique de la Chaleur.* Paris: Firmin Didot Père et Fils, 1822. 639 p. (in French)
5. Friedman A. Asymptotic Behavior of solutions of parabolic equations of any order. *Acta Math.*, 1961. Vol. 106, No. 1–2. P. 1–43. DOI: 10.1007/BF02545812
6. Friedman A. *Partial Differential Equations of Parabolic Type.* New Jersey: Prentice-Hall, Englewood Cliffs, 1964. 347 p.
7. Gilkey P. B. *Invariance Theory, the Heat Equation, and the Atiyah–Singer Index Theorem.* Math. Lect. Ser., vol. 11. Delaware: Publish or Perish, Inc., Wilmington, 1984. 512 p.
8. Il'in A. M., Khas'minskii R. Z. Asymptotic behavior of solutions of parabolic equations and an ergodic property of non-homogeneous diffusion processes. *Math. Sb. (N. S.)*, 1963. Vol. 60, No. 3. P. 366–392. (in Russian)
9. Il'in A. M. *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems.* Transl. Math. Monogr., Vol. 102. Am. Math. Soc., 1992. 281 p.
10. Narasimhan T. N. Fourier's heat conduction equation: History, influence, and connections. *Rev. Geophys.*, 1999. Vol. 37, No. 1. P. 151–172. DOI: 10.1029/1998RG900006
11. Kamin S., Peletier L. A., Vazquez J. L. A nonlinear diffusion-absorption equation with unbounded initial data. In: *Nonlin. Diff. Eq. Equilib. Stat. Progr. Nonlinear Differential Equations Appl., vol. 7. Lloyd N.G., Ni W.M., Peletier L.A., Serrin J. (eds.)* Boston, MA: Birkhäuser, 1992. Vol. 3, P. 243–263. DOI: 10.1007/978-1-4612-0393-3_18
12. Krzyżański M. Sur l'allure asymptotique des solutions d'équations du type parabolique. *Bull. Acad. Polon. Sci.*, 1956. Vol. 4, No. 5. P. 247–251.
13. Lacey A. A., Tzanetis D. E. Global, unbounded solutions to a parabolic equation. *J. Differential Equations*, 1993. Vol. 101, No. 1. P. 80–102. DOI: 10.1006/jdeq.1993.1006
14. Li J. Heat equation in a model matrix geometry. *C. R. Math. Acad. Sci. Paris*, 2015. Vol. 353, No. 4. P. 351–355. DOI: 10.1016/j.crma.2014.10.024
15. Lieberman G. M. *Second Order Parabolic Differential Equations.* River Edge: World Scientific, 1996. 452 p. DOI: 10.1142/3302
16. Poincaré H. Sur les intégrales irrégulières des équations linéaires. *Acta Math.*, 1886. Vol. 8, No. 1. P. 295–344. DOI: 10.1007/BF02417092 (in French)
17. Reynolds A. Asymptotic behavior of solutions of nonlinear parabolic equations. *J. Differential Equations*, 1972. Vol. 12, No. 2. P. 256–261. DOI: 10.1016/0022-0396(72)90032-0
18. Tychonoff A. Théorèmes d'unicité pour l'équation de la chaleur. *Math. Sb.*, 1935. Vol. 42, No. 2. P. 199–216. (in French)
19. Widder D. V. *The Heat Equation.* New York: Academic Press, 1976. 267 p.
20. Zakharov S. V. Heat distribution in an infinite rod. *Math. Notes*, 2006. Vol. 80, No. 3. P. 366–371. DOI: 10.1007/s11006-006-0148-x
21. Zakharov S. V. Two-parameter asymptotics in the Cauchy problem for a quasi-linear parabolic equation. *Asympt. Anal.*, 2009. Vol. 63, No. 1–2. P. 49–54. DOI: 10.3233/ASY-2008-0927

Editor: Tatiana F. Filippova

Managing Editor: Oxana G. Matviychuk

Design: Alexander R. Matviychuk

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990

Phone: +7 (343) 375-34-73

Fax: +7 (343) 374-25-81

Email: secretary@umjuran.ru

Web-site: https://umjuran.ru

N.N. Krasovskii Institute of Mathematics and Mechanics
of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N. Yeltsin

Distributed for free