Springer Proceedings in Mathematics & Statistics

Sandra Pinelas · John R. Graef ·
Stefan Hilger · Peter Kloeden ·
Christos Schinas *Editors*

Differential and Difference Equations with Applications

ICDDEA 2019, Lisbon, Portugal, July 1–5

Springer

Springer Proceedings in Mathematics & Statistics

Volume 333

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Sandra Pinelas · John R. Graef ·
Stefan Hilger · Peter Kloeden ·
Christos Schinas
Editors

Differential and Difference Equations with Applications

ICDDEA 2019, Lisbon, Portugal, July 1–5

Springer

Editors
Sandra Pinelas
Department of Exact Sciences
and Engineering
Military Academy
Lisbon, Portugal

S.M. Nikolskii Mathematical Institute
of RUDN
Moscow, Russia

Stefan Hilger
Mathematik und Didaktik
Katholische Universität Eichstätt
Eichstätt, Bayern, Germany

Christos Schinas
Department of Electrical and Computer
Engineering
Democritus University of Thrace
Xanthi, Xanthi, Greece

John R. Graef
Department of Mathematics
University of Tennessee at Chattanooga
Chattanooga, TN, USA

Peter Kloeden
Huazhong University of Science
and Technology
Wuhan, Hubei, China

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-56322-6 ISBN 978-3-030-56323-3 (eBook)
https://doi.org/10.1007/978-3-030-56323-3

Mathematics Subject Classification: 33C45, 34A33, 34A40, 34B10, 34B15, 34B16, 34B18, 34B37, 34C10, 34C15, 34C37, 34D05, 34H05, 34K06, 34K20, 34K25, 34K26, 34L15, 34K11, 34K45, 34H05, 35B15, 35B40, 35B41, 35L55, 35J20, 35Q30, 35Q55, 35Q80, 37B10, 37E05, 37G10, 37G15, 37L30, 37H99, 37M20, 39A10, 39A13, 39A06, 39A12, 45J05, 47B39, 47B25, 47B37, 47B65, 47H06, 47H10, 47D06, 47J15, 54C60, 65L10, 65M10, 65M20, 65N30, 65N50, 65Q10, 76H05, 76Z05, 92D25, 93B03, 93B40

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

On the Existence of Positive Solutions for the Time-Scale Dynamic Equations on Infinite Intervals

Abdulkadir Dogan

Abstract This paper investigates the existence of positive solutions to time-scale boundary value problems on infinite intervals. By applying the Leggett-Williams fixed point theorem in a cone, some new results for the existence of at least three positive solutions of boundary value problems are found. With infinite intervals, the theorem can be used to prove the existence of solutions of boundary value problems for nonlinear dynamic equations dependence on the delta derivative explicitly. Our results are new for the special cases of difference equations and differential equations as well as in the general time scale setting.

Keywords Fixed point theorems · Time scales · Dynamic equations · Positive solutions · Infinite intervals

MSC 34B15 · 39A10

1 Introduction

The time-scale boundary value problems (BVPs) on infinite intervals arise in a variety of different areas of applied mathematics and physics.

We would like to mention some results of Agarwal and O'Regan [2], Liu [15], Lian and Ge [13], Lian, Pand and Ge [14], Dogan [8].

The book [2] is an excellent source on Infinite Interval Problems for Differential, Difference and Integral Equations. For examples as well as the text [2] by Agarwal and O'Regan for a thorough treatment of the problem

$$y'' + q(t)f(t, y) = 0, \quad 0 < t < +\infty; \quad y(0) = a, \quad \lim_{t \to +\infty} y(t) = 0.$$

A. Dogan (✉)
Department of Applied Mathematics, Faculty of Computer Sciences, Abdullah Gul University, 38039 Kayseri, Turkey
e-mail: abdulkadir.dogan@agu.edu.tr

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_1

In [15], Liu studied the following BVP on the half-line

$$x''(t) + f(t, x(t)) = 0, \qquad t \in (0, +\infty),$$
$$x(0) = 0, \qquad x'(\infty) = y_\infty \geq 0,$$

where $f \in C[(0, +\infty) \times (0, +\infty), [0, +\infty)]$, the author proved that the existence of positive solutions to the above BVP by using a fixed point theorem of cone expansion and compression of norm type.

In [13], Lian and Ge studied the existence of second-order three point BVP on the half line

$$x''(t) + f(t, x(t), x'(t)) = 0, \qquad 0 < t < +\infty,$$
$$x(0) = \alpha x(\eta), \qquad \lim_{t \to +\infty} x'(t) = 0,$$

where $\alpha \in R,\ \alpha \neq 1$ and $\eta \in (0, +\infty)$. They established some criteria for the existence of solutions to the system discussed with suitable conditions imposed on f.

In [14], Lian, Pand and Ge studied the existence of positive solutions for the following BVP with a p-Laplacian operator on a half-line

$$(\varphi_p(x'(t)))' + \phi(t) f(t, x(t), x'(t)) = 0, \qquad 0 < t < +\infty,$$
$$\alpha x(0) - \beta x'(0) = 0, \qquad x'(\infty) = 0.$$

They proved the existence of at least three positive solutions by using a fixed-point theorem in a cone due to Avery-Peterson.

In [8], Dogan studied the following p-Laplacian BVPs on time scales

$$(\phi_p(u^\Delta(t)))^\nabla + a(t) f(t, u(t), u^\Delta(t)) = 0, \qquad t \in [0, T]_{\mathbb{T}},$$
$$u(0) - B_0(u^\Delta(0)) = 0, \qquad u^\Delta(T) = 0,$$

where $\phi_p(u) = |u|^{p-1}u,\ p > 1$. We proved the existence of triple positive solutions for the one-dimensional p-Laplacian BVP by using the Leggett-Williams fixed point theorem.

Motivated by all the works above, we aim to discuss the existence of at least three positive solution of time-scale BVPs on infinite intervals

$$(\varphi_p(x^\Delta(t)))^\nabla + \phi(t) f(x(t), x^\Delta(t)) = 0, \quad t \in (0, \infty)_{\mathbb{T}}, \tag{1.1}$$

$$x(0) - \beta x^\Delta(0) = \gamma x^\Delta(\eta), \qquad \lim_{t \in \mathbb{T},\ t \to \infty} x^\Delta(t) = 0, \tag{1.2}$$

where $\varphi_p(s) = |s|^{p-1}s,\ p > 1, (\varphi_p)^{-1} = \varphi_q, 1/p + 1/q = 1,\ \ \eta \in \mathbb{T}, \eta > 0,\ \ \beta,$ $\gamma \in \mathbb{R},\ \ \beta, \gamma > 0$. Some basic definitions on dynamic equations on time scales can be found in [5, 6, 9].

Throughout this paper, our results assume the following conditions:

(C1) $f \in C([0,\infty)\times[0,\infty),[0,\infty))$ satisfies $f(x,v) \le \omega(\max\{|x|,|v|\})$ with $\omega \in C([0,\infty),[0,\infty))$ nondecreasing;

(C2) $\phi \in C([0,\infty),[0,\infty))$, $\varphi_q\Big(\int_0^\infty \phi(\tau)\nabla\tau\Big) < \infty$, $\int_0^\infty \varphi_q\Big(\int_s^\infty \phi(\tau)\nabla\tau\Big)$ $\Delta s < \infty$;

(C3) $\Upsilon(\delta_1,\delta_2) = \min\limits_{(x,v)\in[\delta_1,\delta_2]\times[0,\delta_2]} f(x,v) > 0, \quad \text{for } 0 < \delta_1 < \delta_2.$

Due to the fact that an infinite interval is noncompact, the discussion about BVPs on the half line is more complicated, in particular, for the time-scale BVPs on infinite intervals. The main methods used on the infinite interval problems are the extension of continuous solutions on the corresponding finite intervals.

Recently, BVPs on time scales for second-order dynamic equations in a finite interval have been extensively studied by many authors [1, 3, 4, 7, 8, 10, 11, 16, 17]. But there is few papers concerned with the existence of positive solutions to the time-scale BVPs of dynamic equations on infinite intervals [18]. To the best knowledge of the author, no one has studied the existence of positive solutions to the time-scale BVP (1.1) and (1.2) by using Leggett-Williams fixed point theorem. Our results of this paper extend and supplement some results from [8, 18].

2 Preliminaries

In this section we present some definitions and lemmas, which will be needed in the proof of the main results.

We consider the space X defined by

$$X = \Big\{x \in C^{\Delta}[0,+\infty)_{\mathbb{T}},\ \sup_{t\in[0,\infty)_{\mathbb{T}}} \mid x(t) \mid < \infty, \quad \lim_{t\in\mathbb{T},t\to+\infty} x^{\Delta}(t) = 0\Big\}$$

with the norm $\|x\| = \max\{\|x\|_1,\ \|x^{\Delta}\|_\infty\}$ where $\|x\|_1 = \sup_{t\in[0,\infty)_{\mathbb{T}}} |x(t)|$, $\|x^{\Delta}\|_\infty = \sup_{t\in[0,\infty)_{\mathbb{T}}} |x^{\Delta}(t)|$. By using the standard arguments, we can find that $(X, \|.\|)$ is a Banach space.

We define the cone $K \subset X$ by

$$K = \Big\{x \in X : x(t) \ge 0, x \text{ is concave and nondecreasing on } [0,+\infty)_{\mathbb{T}}\Big\}.$$

Definition 2.1. Let X be a real Banach space. A nonempty closed convex set $K \subset X$ is called a cone if it satisfies the following conditions

(i) $r_1 u + r_2 v \in K$ for all $u, v \in K$ and all $r_1 \ge 0,\ r_2 \ge 0$,

(ii) $u \in K, \quad -u \in K$ imply $u = 0$.

Every cone $K \subset X$ induces an ordering in X given by $x \le y$ if and only if $y - x \in K$.

Definition 2.2. A map Ψ is said to be a nonnegative continuous concave functional on a cone K of a real Banach space X if $\Psi : K \to [0,\infty)$ is continuous and

$$\Psi(tu + (1-t)v) \geq t\Psi(u) + (1-t)\Psi(v)$$

for all $u, v \in K$ and $t \in [0,1]$. Let $r_1, r_2, r_3 > 0$ be constants,

$$K_{r_3} = \{x \in K : \|x\| < r_3\}, \qquad K(\Psi, r_1, r_2) = \{x \in K : \Psi(x) \geq r_1,\ \ \|x\| \leq r_2\}.$$

Lemma 2.3. *Suppose that (C2) is satisfied.Then the BVP*

$$(\varphi_p(x^{\Delta}(t)))^{\nabla} + \phi(t)f(x(t), x^{\Delta}(t)) = 0, \quad t \in (0,\infty)_{\mathbb{T}}, \tag{2.1}$$

$$x(0) - \beta x^{\Delta}(0) = \gamma x^{\Delta}(\eta), \qquad \lim_{t\in\mathbb{T}, t\to\infty} x^{\Delta}(t) = 0 \tag{2.2}$$

has the unique solution

$$x(t) = \int_0^t \varphi_q\left(\int_s^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right)\Delta s \tag{2.3}$$

$$+\beta\varphi_q\left(\int_0^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right) + \gamma\varphi_q\left(\int_{\eta}^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right). \tag{2.4}$$

Proof. Integrating (2.1) from t to ∞ and using the second condition of (2.2), one gets

$$x^{\Delta}(t) = \varphi_q\left(\int_t^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right). \tag{2.5}$$

Integrating the above equation from 0 to t, we find

$$x(t) = \int_0^t \varphi_q\left(\int_s^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right)\Delta s + x(0). \tag{2.6}$$

Using the first condition of (2.2), we get

$$x(0) - \beta\varphi_q\left(\int_0^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right) = \gamma\varphi_q\left(\int_{\eta}^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right).$$

Hence,

$$x(0) = \beta\varphi_q\left(\int_0^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right) + \gamma\varphi_q\left(\int_{\eta}^{\infty} \phi(\tau)f(x(\tau), x^{\Delta}(\tau))\nabla\tau\right). \tag{2.7}$$

Substituting (2.7) in (2.6), we find

$$x(t) = \int_0^t \varphi_q\left(\int_s^\infty \phi(\tau) f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\Delta s$$
$$+\beta\varphi_q\left(\int_0^\infty \phi(\tau) f(x(\tau), x^\Delta(\tau))\nabla\tau\right) + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau) f(x(\tau), x^\Delta(\tau))\nabla\tau\right).$$

This completes the proof of the lemma. □

To prove our main results, we need the following theorem [12].

Theorem 2.4 *(Leggett-Williams). Let $F : \overline{K}_{r_3} \to \overline{K}_{r_3}$ be a completely continuous map and Ψ be a nonnegative continuous concave functional on K such that $\Psi(u) \leq \|u\|$, $\forall u \in \overline{K}_{r_3}$. Assume that there exist r_1, r_2, r_4 with $0 < r_1 < r_2 < r_4 \leq r_3$ such that*

(C4) $\{u \in K(\Psi, r_2, r_4) : \Psi(u) > r_2\} \neq \emptyset$ *and* $\Psi(Fu) > r_2$, *for all* $u \in K(\Psi, r_2, r_4)$;
(C5) $\|Fu\| < r_1$, *for all* $u \in \overline{K}_{r_1}$;
(C6) $\Psi(Fu) > r_2$, *for all* $u \in K(\Psi, r_2, r_3)$, *with* $\|Fu\| > r_4$.

Then F has at least three fixed points u_1, u_2, u_3 satisfying

$$\|u_1\| < r_1, \quad r_2 < \Psi(u_2), \quad \|u_3\| > r_1, \quad \Psi(u_3) < r_2.$$

3 Main Results

Let the nonnegative continuous concave functional $\Psi : K \to [0, \infty)$ be defined by

$$\Psi(x) = \min_{t\in[\eta,l]_{\mathbb{T}}} |x(t)|, \quad \forall x \in K,$$

where $l \in \mathbb{T}$ be fixed, such that $0 < \eta < l < \infty$. We can easily see that

$$\Psi(x) = x(\eta) \leq \sup_{t\in[0,\infty]_{\mathbb{T}}} |x(t)| \leq ||x||.$$

For convenience, we introduce the following notations. Let

$$\lambda_1 = \varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right), \quad \lambda_2 = (\eta+\beta+\gamma)\varphi_q\left(\int_\eta^l \phi(\tau)\nabla\tau\right),$$
$$\lambda_3 = \int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s + \beta\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right) + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)\nabla\tau\right).$$

Now, we define an operator $A : K \to C[0, +\infty)$ by

$$(Ax)(t) = \int_0^t \varphi_q\left(\int_s^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\Delta s$$
$$+\beta\varphi_q\left(\int_0^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right).$$

The Ascoli-Arzela theorem plays a very important role. But, the Ascoli-Arzela theorem is not suitable for operators on the half line. Therefore, we need a modified compactness criterion to verify A is compact.

Lemma 3.1. *Let (C1) and (C2) hold. Then $A : K \to K$ is completely continuous.*

Proof. We divide the proof in the following four parts.

(1) We claim that $A : K \to K$. Indeed, for all $x \in K$, one has that

$$(Ax)(0) = \beta\varphi_q\left(\int_0^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right)$$
$$+\gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \geq 0,$$
$$(Ax)^\Delta(t) = \varphi_q\left(\int_t^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \geq 0,$$
$$(Ax)^\Delta(\infty) = 0, \quad (\varphi_p((Ax)^\Delta))^\nabla(t) = -\phi(t)f(x(t), x^\Delta(t)) \leq 0.$$

This implies that $A : K \to K$.

(2) We claim that $A : K \to K$ is continuous. Because $f : [0,\infty) \times [0,\infty) \to [0,\infty)$ is continuous, A is continuous. We can readily find this conclusion, so it is omitted here.

(3) We claim that $A : K \to K$ is relatively compact. If Ω is any bounded subset of K, then there exists $M > 0$ such that $||x|| \leq M$ for all $x \in \Omega$. By condition(C2), we obtain

$$(Ax)(t) \leq \omega(M)\int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s$$
$$+\beta\omega(M)\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right) + \gamma\omega(M)\varphi_q\left(\int_\eta^\infty \phi(\tau)\nabla\tau\right) < \infty,$$
$$|(Ax)^\Delta(t)| \leq \omega(M)\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right) < \infty.$$

So $A\Omega$ is uniformly bounded.

Now, we claim that $(A\Omega)^\Delta$ is locally equicontinuous on $[0,\infty)_{\mathbb{T}}$. For any $N > 0$, $t_1, t_2 \in [0, N]_{\mathbb{T}}$ and $x \in \Omega$, without loss of generality we can take that $t_1 < t_2$.

For any $\epsilon > 0$, there is $\delta > 0$ such that if $|t_1 - t_2| < \delta$, then

$$|(\varphi_p((Ax)^\Delta))(t_1) - (\varphi_p((Ax)^\Delta))(t_2)| \leq \omega(M) \int_{t_1}^{t_2} \phi(\tau)\nabla\tau < \epsilon.$$

Thus $(A\Omega)^\Delta$ is equicontinuous on $[0, N]_{\mathbb{T}}$. Because N is arbitrary, $(A\Omega)^\Delta$ is equicontinuous on $[0, \infty)_{\mathbb{T}}$.

(4) We claim that $A : K \to K$ is equiconvergent at ∞. For $\forall x \in \Omega$, from condition (C2), we get

$$\lim_{t\in\mathbb{T}, t\to\infty} |(Ax)(t) - (Ax)(\infty)| \leq \omega(M) \lim_{t\in\mathbb{T}, t\to\infty} \int_t^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s = 0,$$

$$\lim_{t\in\mathbb{T}, t\to\infty} |(\varphi_p((Ax)^\Delta))(t) - (\varphi_p((Ax)^\Delta))(\infty)| \leq \omega(M) \lim_{t\in\mathbb{T}, t\to\infty} \int_t^\infty \phi(\tau)\nabla\tau = 0.$$

So $A\Omega$ is equiconvergent at infinity. Hence, $A : K \to K$ is completely continuous and this proves the lemma.

□

The main result of this paper is following:

Theorem 3.2. *Assume that conditions (C1)–(C3) are satisfied. Suppose that there exist numbers* r_1, r_2, r_4 *such that* $0 < r_1 < r_2 \leq \frac{\lambda_2 \Upsilon(r_2, r_3)}{\lambda_3 \omega(r_3)} r_4 < r_4 \leq r_3$ *and*

(C7) $f(x, v) < \varphi_p(r_1/\lambda_3)$ *for all* $(x, v) \in [0, r_1] \times [0, r_1]$;
(C8) $f(x, v) \leq \varphi_p(r_3/\lambda_3)$ *for all* $(x, v) \in [0, r_3] \times [0, r_3]$;
(C9) $f(x, v) > \varphi_p(r_2/\lambda_2)$ *for all* $(x, v) \in [r_2, r_4] \times [0, r_4]$;
(C10) $\lambda_1 \leq \lambda_3$.

Then BVP (1.1) *and* (1.2) *has at least three positive solutions* x_1, x_2 *and* x_3 *satisfying*

$$||x_1|| < r_1, \quad r_2 < \Psi(x_2), \quad ||x_3|| > r_1 \quad \text{and} \quad \Psi(x_3) < r_2.$$

Proof. The proof is divided into some steps.

(1) We verify that condition (C5) of Theorem 2.4 is satisfied. Assume that there exists a positive number σ such that $f(x, v) \leq \varphi_p(\sigma/\lambda_3)$ for all $(x, v) \in [0, \sigma] \times [0, \sigma]$, then $A\overline{K}_\sigma \subset \overline{K}_\sigma$. Since $A : K \to K$ is completely continuous, we get $A\overline{K}_\sigma \subset K$. Moreover, for all $x \in \overline{K}_\sigma$, we get $0 \leq ||x|| \leq \sigma$. Hence we obtain

$$\begin{aligned}|(Ax)(t)| = \Bigg| &\int_0^t \varphi_q\left(\int_s^\infty \phi(\tau) f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\Delta s \\ &+ \beta\varphi_q\left(\int_0^\infty \phi(\tau) f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \\ &+ \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau) f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\Bigg|\end{aligned}$$

$$\begin{aligned}
&\le \int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\Delta s\\
&\quad+\beta\varphi_q\left(\int_0^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\\
&\quad+\gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\\
&\le \frac{\sigma}{\lambda_3}\Bigg(\int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s+\beta\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right)\\
&\quad+\gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)\nabla\tau\right)\Bigg)=\lambda_3\cdot\frac{\sigma}{\lambda_3}=\sigma.
\end{aligned}$$

In view of assumption (C10), we obtain

$$\begin{aligned}
|(Ax)^\Delta(t)| &= \left|\varphi_q\left(\int_t^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\right|\\
&\le \frac{\sigma}{\lambda_3}\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right)=\lambda_1\cdot\frac{\sigma}{\lambda_3}\le\sigma.
\end{aligned}$$

Therefore $A\overline{K}_\sigma \subset K_\sigma$. Similarly, we can verify that if the conditions (C7) and (C8) are satisfied, then $A\overline{K}_{r_1} \subset K_{r_1}$ and $A\overline{K}_{r_3} \subseteq \overline{K}_{r_3}$.

(2) We verify that condition (C4) of Theorem 2.4 is satisfied. Select $x(t)=\frac{r_2+r_4}{2},\quad 0\le t<+\infty$. It can be checked that the condition (C4) of Theorem 2.4. We can easily see that $x(t)\in K,\quad ||x||=\frac{r_2+r_4}{2}\le r_4,\quad \Psi(x)=\frac{r_2+r_4}{2}>r_2$. We can write

$$\{x\in K(\Psi,r_2,r_4):\Psi(x)>r_2\}\ne\emptyset.$$

In addition, $\forall x\in K(\Psi,r_2,r_4)$, one has $r_2\le x(t)\le r_4,\quad$ for $\quad t\in[\eta,l],\quad ||x||\le r_4$. From condition (C9), we find

$$\begin{aligned}
\Psi(Ax)=(Ax)(\eta)&=\int_0^\eta \varphi_q\left(\int_s^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\Delta s\\
&\quad+\beta\varphi_q\left(\int_0^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\\
&\quad+\gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\\
&\ge \eta\varphi_q\left(\int_\eta^l \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)+\beta\varphi_q\left(\int_\eta^l \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\\
&\quad+\gamma\varphi_q\left(\int_\eta^l \phi(\tau)f(x(\tau),x^\Delta(\tau))\nabla\tau\right)\\
&>\frac{r_2}{\lambda_2}\Big(\eta+\beta+\gamma\Big)\varphi_q\left(\int_\eta^l \phi(\tau)\nabla\tau\right)=r_2.
\end{aligned}$$

(3) We verify that condition (C6) of Theorem 2.4 is satisfied. For $\forall x \in K(\Psi, r_2, r_4)$, and $||Ax|| > r_4$, one has $r_2 \le x(t) \le r_3$, for $t \in [\eta, l]$, $||x|| \le r_3$. By conditions (C3) and (C10), we obtain

$$\begin{aligned}
\Psi(Ax) &= (Ax)(\eta) = \int_0^\eta \varphi_q\left(\int_s^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\Delta s \\
&\quad + \beta\varphi_q\left(\int_0^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \\
&\quad + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \\
&\ge \eta\varphi_q\left(\int_\eta^l \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) + \beta\varphi_q\left(\int_\eta^l \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \\
&\quad + \gamma\varphi_q\left(\int_\eta^l \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \\
&\ge \Upsilon(r_2, r_3)(\eta + \beta + \gamma)\varphi_q\left(\int_\eta^l \phi(\tau)\nabla\tau\right) \\
&= \frac{\Upsilon(r_2, r_3)(\eta + \beta + \gamma)\varphi_q\left(\int_\eta^l \phi(\tau)\nabla\tau\right)\omega(r_3)}{\omega(r_3)\Big(\int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s + \beta\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right) + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)\nabla\tau\right)\Big)} \\
&\quad \times\left(\int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s + \beta\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right)\right. \\
&\quad \left. + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)\nabla\tau\right)\right) \\
&= \frac{\lambda_2\Upsilon(r_2, r_3)}{\lambda_3\omega(r_3)}\omega(r_3) \times \left(\int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)\nabla\tau\right)\Delta s\right. \\
&\quad \left. + \beta\varphi_q\left(\int_0^\infty \phi(\tau)\nabla\tau\right) + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)\nabla\tau\right)\right) \\
&\ge \frac{\lambda_2\Upsilon(r_2, r_3)}{\lambda_3\omega(r_3)}\left(\int_0^\infty \varphi_q\left(\int_s^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\Delta s\right. \\
&\quad + \beta\varphi_q\left(\int_0^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right) \\
&\quad \left. + \gamma\varphi_q\left(\int_\eta^\infty \phi(\tau)f(x(\tau), x^\Delta(\tau))\nabla\tau\right)\right) \\
&\ge \frac{\lambda_2\Upsilon(r_2, r_3)}{\lambda_3\omega(r_3)}||Ax|| > \frac{\lambda_2\Upsilon(r_2, r_3)}{\lambda_3\omega(r_3)}r_4 \ge r_2.
\end{aligned}$$

Hence, by Theorem 2.4, we know that BVP (1.1) and (1.2) has at least three positive solutions x_1, x_2 and x_3 such that

$$||x_1|| < r_1, \quad r_2 < \Psi(x_2), \quad ||x_3|| > r_1 \quad \text{and} \quad \Psi(x_3) < r_2.$$

This completes the proof of the theorem. □

Acknowledgments The author gratefully acknowledges Application Number 1919B021900156 of 2224-International Scientific Meetings Fellowship Programme 2019/1 by the Scientific and Technological Research Council of Turkey (TÜBİTAK).

References

1. Agarwal, R.P., O'Regan, D.: Nonlinear boundary value problems on time scales. Nonlinear Anal. **44**, 527–535 (2001)
2. Agarwal, R.P., O'Regan, D.: Infinite Interval Problems for Differential, Difference and Integral Equations. Kluwer Academic Publisher, Dordrecht (2001)
3. Anderson, D.R., Zhai, C.: Positive solutions to semi-positone second-order three-point problems on time scales. Appl. Math. Comput. **215**, 3713–3720 (2010)
4. Anderson, D., Avery, R., Henderson, J.: Existence of solutions for a one dimensional p-Laplacian on time-scales. J. Diff. Equ. Appl. **10**, 889–896 (2004)
5. Bohner, M., Peterson, A.: Dynamic Equations on Time Scales: An Introduction with Applications. Birkhauser, Boston (2001)
6. Bohner, M., Peterson, A.: Advances in Dynamic Equations on Time Scales. Birkhauser, Boston (2003)
7. DaCunha, J.J., Davis, J.M., Singh, P.K.: Existence results for singular three point boundary value problems on time scales. J. Math. Anal. Appl. **295**, 378–391 (2004)
8. Dogan, A.: On the existence of positive solutions for the one-dimensional p-Laplacian boundary value problems on time scales. Dyn. Syst. Appl. **24**, 295–304 (2015)
9. Georgiev, S.: Integral Equations on Time Scales. Atlantis Press, Paris (2016)
10. Goodrich, C.S.: The existence of a positive solution to a second-order delta-nabla p-Laplacian BVP on a time scale. Appl. Math. Lett. **25**, 157–162 (2012)
11. He, Z., Li, L.: Multiple positive solutions for the one-dimensional p-Laplacian dynamic equations on time scales. Math. Comput. Modell. **45**, 68–79 (2007)
12. Leggett, R.W., Williams, L.R.: Multiple positive fixed points of nonlinear operators on ordered Banach spaces. Indiana Univ. Math. J. **28**, 673–688 (1979)
13. Lian, H., Ge, W.: Solvability for second-order three-point boundary value problems on a half-line. Appl. Math. Lett. **19**, 1000–1006 (2006)
14. Lian, H., Pang, H., Ge, W.: Triple positive solutions for boundary value problems on infinite intervals. Nonlinear Anal. **67**, 2199–2207 (2007)
15. Liu, Y.: Existence and unboundedness of positive solutions for singular boundary value problems on half-line. Appl. Math. Comput. **144**, 543–556 (2003)
16. Sun, H.R., Tang, L.T., Wang, Y.H.: Eigenvalue problem for p-Laplacian three-point boundary value problems on time scales. J. Math. Anal. Appl. **331**, 248–262 (2007)
17. Wang, D.B.: Three positive solutions of three-point boundary value problems for p-Laplacian dynamic equations on time scales. Nonlinear Anal. **68**, 2172–2180 (2008)
18. Zhao, X., Ge, W.: Multiple positive solutions for time scale boundary value problems on infinite intervals. Acta Appl. Math. **106**, 265–273 (2009)

A Randomized Quasi-Monte Carlo Algorithms for Some Boundary Value Problems

Alexander S. Sipin

Abstract This work continues the study of stochastic algorithms for solving boundary value problems, which started in our previous papers. The Dirichlet problem for the Laplace equation are discussed. We compare Monte Carlo and randomized quasi-Monte Carlo versions of algorithms. We use the Halton random points constructed by the Cranley-Patterson method.

1 Introduction

For the numerical solution of boundary value problems, various numerical methods are used, including statistical modeling methods, i.e. Monte Carlo methods (see, for example, [1, 2]). Effective statistical modeling procedures have been developed to solve the equations of radiation transfer, gas dynamics equations, a number of problems in the field of electrostatics, elasticity theory and others. Statistical algorithms allow solving boundary value problems both inside and outside a bounded domain, the boundary of which can have a complex structure. For a wide class of problems, computational work in such algorithms linearly depends on the dimension of the domain.

When a statistical algorithm is constructing, the solution of a boundary value problem is written in the form of a mathematical expectation of some random variable ξ. That is, the random variable ξ is an unbiased estimator of the solution of the boundary value problem. Usually unbiased estimators for solving boundary value problems are constructed on the trajectories of random walks. We use a random walk on spheres to solve the Dirichlet problem. To simplify the formulas, we consider only three-dimensional problem for the Laplace equation. Thus, the paper considers algorithms for calculating the value of a harmonic function $u(x)$ at point x of a three-dimensional bounded domain from the known values of this function in boundary currents. Any simulating procedure for the estimator ξ can be written as a function of a sequence of independent random variables distributed uniformly over a segment [0, 1]. Therefore, the solution $u(x)$ can be written as a sum of integrals over some

A. S. Sipin (✉)
Vologda State University, Lenina, 15, Vologda, Russia
e-mail: cac1909@mail.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_2

s-dimensional unit cube $[0, 1]^s$. Dimension s can reach several hundred. To calculate such integrals, it is recommended to use the quasi- Monte Carlo method, which is more efficient than the Monte Carlo method with unlimited increase in sample size. A numerical comparison of Monte Carlo and quasi-Monte Carlo versions of this algorithm can be found in [3]. It has been shown that the real benefits of the quasi-Monte Carlo method begin with sample sizes exceeding 10^7. The comparison was carried out with known exact solutions of boundary value problems, since for the quasi Monte Carlo method it is impossible to estimate the error in the course of calculations. In this paper we compare numerically Monte Carlo and randomized quasi-Monte Carlo version of random walk on spheres algorithm. We compare the statistical errors of these algorithms and determine the sample size at which the randomized quasi Monte Carlo method becomes more profitable.

2 Application of the Mean Value Theorem for a Harmonic Function to Calculate its Values

Let $u(x)$ be a harmonic function in a bounded domain $\mathscr{D} \subset R^3$ and let $u(x)$ be continuous in $\overline{\mathscr{D}}$. The distance from the point $x \in \mathscr{D}$ to the boundary Γ we denote by $d(x)$. Let $\omega(1), \omega(2), ...$ be a sequence independent random vectors uniformly distributed on a sphere of radius 1 centered at zero.

For any point $x \in \mathscr{D}$, by the mean value theorem, we obtain

$$u(x) = Eu(x + d(x)\omega(1)), \tag{1}$$

where E is a symbol of mathematical expectation of a random variable. Let

$$x(0) = x, \quad x(k+1) = x(k) + d(x(k))\omega(k+1), \quad k = 0, 1, 2, ..., \tag{2}$$

then after m iterations of the formula (1) we have

$$u(x) = Eu(x(m)). \tag{3}$$

The random process defined by formula (2) is called Random Walk on Spheres.

Assuming the function $u(x)$ is known, we will use $\xi = u(x(m))$ as an unbiased estimator for $u(x)$. To get the value of ξ, you need to simulate the sequence $\omega(1), \omega(2), ..., \omega(m)$. We get it using standard formulas for modeling an isotropic unit vector

$$\begin{aligned} \omega_1(i) &= 2\alpha_{2i-1} - 1, \\ \omega_2(i) &= \sqrt{1 - \omega_1^2(i)}\cos(2\pi\alpha_{2i}), \\ \omega_3(i) &= \sqrt{1 - \omega_1^2(i)}\sin(2\pi\alpha_{2i}), \end{aligned} \tag{4}$$

where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{2m})$ is a point uniformly distributed in the $2m$-dimensional unit cube $\overline{I}^{2m} = [0; 1]^{2m}$.

The mathematical expectation any function $f(\alpha)$ coincides with her integral over the hypercube, therefore we obtain a representation $u(x)$ in the form of an integral over hypercube $\overline{I}^{2m}$. We use this integral to compare the efficiency of Monte Carlo, quasi-Monte Carlo and randomized quasi-Monte Carlo methods. The test is quite complicated, especially when the dimension m is several hundred.

We calculate this integral by Monte Carlo, using pseudo-random numbers, which are generated using multiplicative congruential method. In this case, we denote the points inside the hypercube by $\alpha(n)$ and call them pseudo-random points. The integral is calculated as the average value of $u(x^{(n)}(m))$ over a large number N trajectories of Random Walk on Spheres

$$u(x) \approx \frac{1}{N} \sum_{n=1}^{N} u(x^{(n)}(m)), \tag{5}$$

where $x^{(n)}(m)$ is the last point for $n-$th trajectory of the Random Walk on Spheres process.

In the case of quasi-Monte Carlo, to simulate the process, we use non-random Halton points $\alpha^H(n)$, which for any integer $n \geq 0$ are defined by formulas

$$\alpha^H(n) = (\phi_{b_1}(n), \phi_{b_2}(n), ...\phi_{b_{2m}}(n)), \tag{6}$$

where $\phi_{b_i}(n)$ for $1 \leq i \leq 2m$ is the radical-inverse function [5] in base b_i and b_i is the element number i in a sequence of primes 2, 3, 5, 7, ... To calculate the integral, the formula (5) is again used. The performance of the Monte Carlo and Quasi-Monte Carlo methods in this test is discussed in our previous work [3].

We use Cranley-Patterson's rotation (see [4]) to construct a randomized quasi-Monte Carlo method. So, we use random Halton points $\alpha_t^H(n) \in \overline{I}^{2m}, t = 1, 2 \ldots, T$, to simulate the process. They are calculate by formula

$$\alpha_t^H(n) = (\alpha^H(n) + \alpha(t)) \quad mod \quad 1, \tag{7}$$

where function $x \ mod \ 1$ is the fractional part of x and $\alpha(t) \in \overline{I}^{2m}$, $t = 1, \ldots, T$ are independent random points (really pseudorandom points). Now we can use the estimator

$$\mu = \frac{1}{T} \sum_{t=1}^{T} \mu_t = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{n=1}^{N} u(x_t^{(n)}(m)) \right), \tag{8}$$

where $x_t^{(n)}(m)$ is the last point for $n-$th trajectory of the Random Walk on Spheres process constructed using random Holton points. The random variables μ_t and μ are an unbiased estimators for $u(x)$ and variance of random variable μ_t is constant σ^2.

Therefore, we have

$$Var(\mu) = \frac{\sigma^2}{T} = E\widehat{S}^2 = E\left(\frac{1}{T\cdot(T-1)}\sum_{t=1}^{T}(\mu_t-\mu)^2\right). \tag{9}$$

We use $3\widehat{S}$ as randomized quasi-Monte Carlo accuracy.

Monte Carlo accuracy defined by asymptotic 99,7% confidence interval

$$\left|\int_{\overline{I}^{2m}} g(\alpha)d\alpha - \frac{1}{N}\sum_{n=1}^{N} g(\alpha(n))\right| \leq \frac{3S}{\sqrt{N}}, \tag{10}$$

where S^2 is sample variance.

Quasy-Monte Carlo accuracy for smooth function defined by Koksma-Hlawka inequality [5]. For the Halton points it has the form

$$\left|\int_{\overline{I}^{2m}} g(\alpha)d\alpha - \frac{1}{N}\sum_{n=1}^{N} g(\alpha^H(n))\right| \leq c(2m)\frac{(ln(N))^{2m}}{N}V(g), \tag{11}$$

where $V(g)$ is the variation of the function g in the sense of Hardy and Krause. The variation $V(g)$ is harder to compute than integral itself.

2.1 Numerical Results

The results of calculations by formula (8) when $T = 10$ for two harmonic functions in a cube $[0, 10]^3$ are given in Tables 1, 2. Table 1 shows the results of the deviation of the found approximate values of the harmonic function $u = 3x_1x_2^2 - x_1^3 + 8x_3$ from its exact value $u(x) = 205.472$ at the point $x = (1.6, 5.4, 8.7)$. For $m = 1$, the randomezed quasi-Monte Carlo method is better than Monte Carlo. It is seen that starting from the dimension of the integral 20 ($m = 10$) and the sample size $N <= 10^6$ Monte Carlo method and randomezed quasi-Monte Carlo method gives similar results.

The Table 2 shows, that for the value $u(x) = 1.714986$ of the function $u = 1/\sqrt{(x_1+0.1)^2 + x_2^2 + x_3^2}$ at the point $x = (0.2, 0.4, 0.3)$, both methods gives similar results for $m \geq 10$ and $N \leq 10^6$.

Table 1 Deviation of Mean Value Operator iterations for the function $u = 3x_1x_2^2 - x_1^3 + 8x_3$ from the exact value at the point $x = (1.6, 5.4, 8.7)$. Monte Carlo (MC) and randomized quasi-Monte Carlo (RQM) methods.

m\N	10^4 MC	10^4 RQM	Err MC	Err RQM	10^6 MC	10^6 RQM	Err MC	Err RQM
1	−0.205	−0.005	0.696	0.010	−0.011	0	0.069	8e-5
10	0.643	−0.142	2.110	1.493	0.010	0.002	0.211	0.071
50	0.405	2.036	2.598	2.861	0.373	0.003	0.259	0.176
100	0.178	1.151	2.591	2.183	0.383	0.021	0.260	0.189
150	−0.329	0.552	2.592	1.757	0.363	0.051	0.260	0.115
200	1.122	−0.267	2.605	2.340	0.450	−0.091	0.260	0.247
250	−0.161	0.809	2.580	1.056	0.431	0.026	0.260	0.156

Table 2 Deviation of Mean Value Operator iterations for the function $u = 1/\sqrt{(x_1+0.1)^2 + x_2^2 + x_3^2}$ from the exact value at the point $x = (0.2, 0.4, 0.3)$. Monte Carlo (MC) and randomized quasi-Monte Carlo (RQM) methods.

m\N	10^4 MC	10^4 RQM	Err MC	Err RQM	10^6 MC	10^6 RQM	Err MC	Err RQM
1	−0.0003	0.0000	0.0033	0.0001	0.0001	0.0000	0.0003	4e-7
10	0.0016	0.0002	0.0062	0.0032	0.0001	0.0000	0.0006	0.0001
50	0.0025	−0.0028	0.0068	0.0047	0.0002	0.0000	0.0007	0.0005
100	0.0014	0.0011	0.0068	0.0075	−0.0001	−0.0001	0.0007	0.0003
150	0.0045	0.0016	0.0067	0.0034	0.0005	−0.0001	0.0007	0.0005
200	0.0019	−0.0023	0.0067	0.0065	0.0006	−0.0004	0.0007	0.0004
250	−0.0011	0.0014	0.0067	0.0043	0.0001	0.0000	0.0007	0.0004

3 The Dirichlet Problem for Harmonic Function

Now, we briefly describe the Random Walk on Spheres algorithm for solving the Dirichlet problem for the Laplace equation in the domain $\mathscr{D}$. For any $\varepsilon > 0$, Γ_ε denote an $\varepsilon-$ neighborhood of the boundary Γ. Let $\tau = \min(k : x(k) \in \Gamma_\varepsilon)$ be the first hitting time of the process $x(k)$ into Γ_ε. The Random Walk on Spheres $x(k)$ converges to the boundary Γ with probability 1, hence $\tau < +\infty$ with probability 1. Then τ is a Markov moment for the process $x(k)$ and the equality $u(x) = Eu(x(\tau))$ is true. Hence, we can use the formula

$$u(x) \approx \frac{1}{NT}\sum_{n=1}^{NT} u(x^{(n)}(\tau)) \tag{12}$$

to calculate $u(x)$. Here $x^{(n)}(\tau)$ is the first point lying in Γ_ε for $n-$th trajectory of the Random Walk on Spheres.

Table 3 Random Walk on Spheres. Deviation of Mean Value for the function $u = 1/\sqrt{(x_1+0.1)^2 + x_2^2 + x_3^2}$ from the exact value $u(x) = 0.41922$ at the point $x = (1.1, 0.5, 2)$. Monte Carlo (MC) and randomized quasi-Monte Carlo (RQM) methods.

N	MC	ErrMC	RQM	ErrRQM	L	QL	Lmax	QLmax
10^3	−0.0023	0.0046	0.0006	0.0046	22	22	154	110
10^4	−0.0008	0.0015	0.0001	0.0013	22	21	136	138
10^5	0.00002	0.00048	−0.00003	0.00018	22	22	209	176
10^6	−0.00007	0.00015	−0.00004	0.00008	22	22	196	197

We use the estimator

$$u(x) \approx \frac{1}{T}\sum_{t=1}^{T}\mu_t = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{n=1}^{N} u(x_t^{(n)}(\tau))\right), \tag{13}$$

where $x_t^{(n)}(\tau)$ is the first point lying in Γ_ε for $n-$th trajectory of the Random Walk on Spheres process constructed using randomized Holton points.

3.1 Numerical Results for T = 10

For testing the algorithms, we use the harmonic functions from paper [3] to be able to compare the results for a randomized quasi-Monte Carlo from this paper with the results for quasi-Monte Carlo [3]. The results of calculations by formulas (12), (13) for two harmonic functions in a cube $[0, 10]^3$ are given in Tables 3, 4. In both examples the parameter ε has been chosen equal to 0.001. In the randomized quasi-Monte Carlo method, the value $\tau_m = \min(\tau, m) \quad (m = 1500)$ is used instead of τ.

Variables L and QL denote the average length of the random walk trajectory. The $Lmax$ and $QLmax$ variables denote the maximum length of the random walk trajectory. Again, the results of Monte Carlo and randomized quasi-Monte Carlo are similar when $N \leq 10^6$. We also see that the deviation of the average value from the exact value does not exceed the statistical error of both Monte Carlo and randomized quasi-Monte Carlo.

Let's compare the accuracy of the calculations for quasi-Monte Carlo and randomized quasi Monte Carlo, using results [3], presented in Tables 5, 6. Now the column % shows the relative error of the calculated approximate value of the function. It was used to compare the accuracy of calculations in the Monte Carlo and Quasi-Monte Carlo methods (see [3]). Analyzing the data in the corresponding rows of Tables 3 and 5, as well as Tables 4 and 6, we see that the randomization of quasi-Monte Carlo does not significantly affect the accuracy of the calculations.

Table 4 Random Walk on Spheres. Deviation of Mean Value for the function $u = 3x_1x_2^2 - x_1^3 + 8x_3$ from the exact value $u(x) = 478$ at the point $x = (3, 7, 8)$. Monte Carlo (MC) and randomized quasi-Monte Carlo (RQM) methods.

N	MC	ErrMC	RQM	ErrRQM	L	QL	Lmax	QLmax
10^3	−3.72	14.57	4.60	14.77	25	24	123	114
10^4	−0.96	4.59	0,08	4.06	25	25	161	165
10^5	0,04	1.45	−0,28	0.76	25	25	178	191
10^6	−0,07	0.46	0,14	0.27	25	25	220	220

Table 5 Random Walk on Spheres. Deviation of Mean Value for the function $u = 1/\sqrt{(x_1 + 0.1)^2 + x_2^2 + x_3^2}$ from the exact value $u(x) = 0.41922$ at the point $x = (1.1, 0.5, 2)$. Monte Carlo (MC) and quasi-Monte Carlo (QMC) methods.

N	MC	%	QMC	%	L	QL	Lmax	QLmax
10^4	0.00196	0.47	−0.00376	0.90	22	23	108	105
10^5	−0.00024	0.06	−0.00120	0.29	22	22	140	194
10^6	−0.00003	0.01	−0.00021	0.05	22	22	201	210
10^7	−0.00005	0.01	0.00001	0.002	22	22	201	214

Table 6 Random Walk on Spheres. Deviation of Mean Value for the function $u = 3x_1x_2^2 - x_1^3 + 8x_3$ from the exact value $u(x) = 478$ at the point $x = (3, 7, 8)$. Monte Carlo (MC) and quasi-Monte Carlo (QMC) methods.

N	MC	%	QMC	%	L	QL	Lmax	QLmax
10^4	0,28413	0.06	13,90295	2.91	25	24	122	1257
10^5	−0.00024	0.29	2,75519	0.58	25	25	160	1257
10^6	−0,17767	0.04	0,50839	0.11	25	25	178	1257
10^7	0,09103	0.02	0,12939	0.03	25	25	262	1257

3.2 Numerical Results for T = 100

Now we study the influence of the parameter T on the statistical error in the quasi-Monte Carlo method. The results of the calculations are in Tables 7 and 8. A comparison of the ErrRQM column in Tables 3 and 7 and Tables 4 and 8 shows that the statistical error of the RQM method increases with increasing T. This is due to a decrease in N, which leads to an increase in the error of the QMC algorithm. We also note the practical equality of the values ErrMC and ErrRQM both in Table 7 and in Table 8.

Table 7 Random Walk on Spheres. Deviation of Mean Value for the function $u = 1/\sqrt{(x_1 + 0.1)^2 + x_2^2 + x_3^2}$ from the exact value $u(x) = 0.41922$ at the point $x = (1.1, 0.5, 2)$. Monte Carlo (MC) and randomized quasi-Monte Carlo (RQM) methods.

N	MC	ErrMC	RQM	ErrRQM	L	QL	Lmax	QLmax
10^2	0.003	0.0048	0.001	0.0049	22	22	116	117
10^3	−0.00003	0.0015	0.0005	0.0015	22	22	130	135
10^4	−0.00002	0.00048	−0.00002	0.00042	22	22	215	221
10^5	−0.00015	0.00015	0.000001	0.00011	22	22	234	217

Table 8 Random Walk on Spheres. Deviation of Mean Value for the function $u = 3x_1x_2^2 - x_1^3 + 8x_3$ from the exact value $u(x) = 478$ at the point $x = (3, 7, 8)$. Monte Carlo (MC) and randomized quasi-Monte Carlo (RQM) methods.

N	MC	ErrMC	RQM	ErrRQM	L	QL	Lmax	QLmax
10^2	2.97	14.53	4.60	13.58	24	25	128	125
10^3	0.57	4.59	1.60	4.06	25	25	177	145
10^4	−0,54	1.45	0,23	1.20	25	25	186	172
10^5	−0,11	0.46	0,04	0.29	25	25	213	259

4 Conclusion

The results of computational experiments allow us to draw the following conclusion:

1. The randomized quasi-Monte Carlo method can be used to solve Dirichlet boundary value problem.
2. For a fixed product TN, an increase in the parameter T impairs the accuracy of the RQM method, when $TN \leq 10^7$.
3. The statistical errors of the Monte Carlo and randomized quasi-Monte Carlo methods have the same order of smallness,when $TN \leq 10^7$.
4. The statistical error of the randomized quasi-Monte Carlo method can be used to evaluate the accuracy of the method.
5. When solving a boundary value problem, the randomized quasi-Monte Carlo method has no advantages over the Monte Carlo method if the sample size is $TN \leq 10^7$.

Acknowledgements The research has been supported by the Russian Foundation for Basic Research, project No. 17-01-00267.

References

1. Sabelfeld, K.K.: Monte Carlo Methods in Boundary Value Problems. Springer, Heidelberg (1991)
2. Ermakov, S.M., Nekrutkin, V.V., Sipin, A.S.: Random Processes for Classical Equations of Mathematical Physics. Kluwer Academic Publishers, Dordrecht/Boston/London (1989)
3. Sipin, A.S., Zeifman, A.I.: Numerical experiments for some Markov models for solving boundary value problems. In: Dimov, I., Farago, I., Vulkov, L. (eds.) Finite Difference Methods. Theory and Applications. FDM 2018. Lecture Notes in Computer Science, vol. 11386, pp. 493–500. Springer, Cham (2018)
4. Owen, A.B.: A randomized Halton algorithm in R, tech. report, Stanford University. arXiv:1706.02808 (2017)
5. Niederreiter, H.: Random number generation and quasi-Monte Carlo methods. In: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 63. Society for Industrial and Applied Mathematics, Philadelphia (1992)

On the Study of Forward Kolmogorov System and the Corresponding Problems for Inhomogeneous Continuous-Time Markov Chains

Alexander Zeifman

Abstract An inhomogeneous continuous-time Markov chain $X(t)$ with finite or countable state space under some natural additional assumptions is considered. As a consequence, we study a number of problems for the corresponding forward Kolmogorov system, which is the linear system of differential equations with special structure of the matrix $A(t)$. In the countable situation we have an equation in the space of sequences l_1. The important properties of $X(t)$ (such as weak and strong ergodicity, perturbation bounds, truncation bounds) are closely connected with behaviour of the solutions of the forward Kolmogorov system as $t \to \infty$. The main problems and some approaches for their solution are discussed in the paper.

Keywords Forward Kolmogorov system · Markov chains

1 Introduction

Continuous-time Markov chains are widely used for the study of stochastic models in the natural and technical sciences, such as queuing theory, biology, chemistry, etc.

Let $\{X(t),\ t \geq 0\}$ be a continuous-time Markov chain with state space $\mathscr{X} = \{0, 1, 2\dots\}$. Denote by $p_{ij}(s,t) = P\{X(t) = j\,|X(s) = i\},\ i, j \geq 0,\ 0 \leq s \leq t$ the transition probabilities of $X(t)$ and by $p_i(t) = P\{X(t) = i\}$ – the probability that the Markov chain $X(t)$ is in state i at time t. Let $\mathbf{p}(t) = (p_0(t), p_1(t), \dots)^T$ be the probability distribution vector at instant t. Throughout the paper we assume that in an element of time h the possible transitions and their associated probabilities are

$$p_{ij}(t, t+h) = \begin{cases} q_{ij}(t)h + \alpha_{ij}\,(t,h)\,, & \text{if } j \neq i \\ 1 + q_{ii}(t)h + \alpha_i\,(t,h)\,, & \text{if } j = i, \end{cases} \quad i, j \in \mathscr{X}, \tag{1}$$

A. Zeifman (✉)
Institute of Informatics Problems FRC CSC RAS, Vologda Research Center RAS, Vologda State University, Lenina, 15, Vologda, Russia
e-mail: a_zeifman@mail.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_3

where all the $\alpha_i(t, h)$ are $o(h)$ uniformly in i, i.e. $\sup_i |\alpha_i(t, h)| = o(h)$ and

$$q_{ii}(t) = - \sum_{k \in \mathscr{X}, k \neq i} q_{ik}(t).$$

The matrix $Q(t) = (q_{ij}(t))_{i,j=0}^{\infty}$ is called the intensity (or infinitesimal) matrix of the chain $\{X(t),\ t \geq 0\}$.

The Markov chain $X(t)$ is called homogeneous if Q is a constant matrix, and it is called inhomogeneous in the opposite case.

As a rule, in the inhomogeneous case we will assume that the intensity functions $q_{ij}(t)$ are locally integrable on the interval $[0, \infty)$.

Henceforth it is assumed that the $Q(t)$ is essentially bounded, i.e.

$$\sup_i |q_{ii}(t)| = L(t) \leq L < \infty, \tag{2}$$

for almost all $t \geq 0$.

In many problems, condition (2) can be weakened and replaced by $L(t) < \infty$, for almost all $t \geq 0$.

Then the probabilistic dynamics of the process $\{X(t),\ t \geq 0\}$ is given by the forward Kolmogorov system

$$\frac{d}{dt}\mathbf{p}(t) = A(t)\mathbf{p}(t), \tag{3}$$

where $A(t) = Q^T(t)$ is the transposed intensity matrix. All column sums of this matrix are zeros for any $t \geq 0$, and $A(t)$ is essentially nonnegative (i.e. all its off-diagonal elements are nonnegative for any $t \geq 0$).

Throughout the paper by $\|\cdot\|$ we denote the l_1-norm, i. e. $\|\mathbf{p}(t)\| = \sum_{k \in \mathscr{X}} |p_k(t)|$, and $\|Q(t)\| = \sup_{j \in \mathscr{X}} \sum_{i \in \mathscr{X}} |q_{ij}|$. Let Σ be a set all stochastic vectors, i. e. l_1 vectors with non-negative coordinates and unit norm. Hence we have $\|A(t)\| = 2 \sup_{k \in \mathscr{X}} |q_{kk}(t)| \leq 2L$ for almost all $t \geq 0$. Hence the operator function $A(t)$ from l_1 into itself is bounded for almost all $t \geq 0$ and locally integrable on $[0; \infty)$. Therefore we can consider (3) as a differential equation in the space l_1 with bounded operator.

It is well known (see [2]) that the Cauchy problem for differential Eq. (3) has a unique solutions for an arbitrary initial condition, and $\mathbf{p}(s) \in \Sigma$ implies $\mathbf{p}(t) \in \Sigma$ for $t \geq s \geq 0$.

Denote by $E(t, k) = E(X(t)|X(0) = k)$ the conditional expected number of 'particles' in the system at instant t, provided that initially (at instant $t = 0$) k 'particles were present in the system.

In order to obtain perturbation bounds we consider a class of perturbed Markov chains $\{\bar{X}(t), t \geq 0\}$ defined on the same state space $\mathscr{X}$ as the original Markov chain $\{X(t), t \geq 0\}$, with the intensity matrix $\bar{A}(t)$ and the same restrictions as imposed on $A(t)$. It is assumed that $\|\hat{A}(t)\| = \|A(t) - \bar{A}(t)\| \leq \varepsilon$, for almost all $t \geq 0$, which means the perturbations are considered to be small.

Before proceeding to the derivation of the main results of the paper, we recall two definitions. Recall that a Markov chain $\{X(t),\ t \geq 0\}$ is called *weakly ergodic*, if $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \to 0$ as $t \to \infty$ for any initial conditions $\mathbf{p}^*(0)$ and $\mathbf{p}^{**}(0)$, where $\mathbf{p}^*(t)$ and $\mathbf{p}^{**}(t)$ are the corresponding solutions of (3). A Markov chain $\{X(t),\ t \geq 0\}$ has the limiting mean $\varphi(t)$, if $\lim_{t\to\infty} (\varphi(t) - E(t, k)) = 0$ for any k.

It is clear, therefore, that the study of the qualitative properties and the derivation of estimates for Markov chains with continuous time is reduced to the study of the corresponding properties of solutions of the forward Kolmogorov (3) system on Σ.

A general approach to obtaining sharp bounds on the rate of convergence via the notion of the logarithmic norm of an operator function was recently discussed in detail in our papers [34–36]. The first studies in this direction were published since 1980-s for birth-death models, see [25, 26]. In [34, 35] we have highlighted four fairly broad classes of finite and countable Markov chains, for which the forward Kolmogorov system can be transformed into a system with an essentially nonnegative matrix. Moreover, it turns out that similar results can be obtained for some other models, see, for example [37]. Computation of the limiting characteristics for such chains using bounds on the rate of convergence and truncations technique introduced in [30, 33].

The approach is based on studying the norm of the Cauchy operator of the reduced forward Kolmogorov system by estimation of the so-called logarithmic norm of an operator function. The method of the complete study of the process $X(t)$ that describes the number of claims in the system assumes the construction of a) upper bounds for the rate of convergence of the limit mode, providing that, beginning from a certain time, say, t^*, the probability characteristics of the process $X(t)$ do not depend on the initial conditions (up to a given discrepancy); b) analogous lower bounds which are also very important and provide that the “independence” of the initial conditions cannot appear before a certain time, say, t_*; c) stability bounds providing that if the structure of the matrix of intensities of the process is taken into account in an appropriate way, and the errors in intensities are small, then the basic characteristics of the process are calculated in an adequate way; d) approximations to the process by means of truncation by similar processes with a lesser number of states and construction of the corresponding estimates for the error. Finally, applying the results of a), c), d) to the system with 1-time-periodic intensities and solving the forward Kolmogorov system with the simplest initial condition $X(0) = 0$ for the truncated process on the interval $[t^*, t^* + 1]$, as a result we obtain all basic probability characteristics of both the process $X(t)$, and close “perturbed” processes. Note that the item a) is most important, because after the corresponding bounds are obtained, the solutions of other problems can be constructed automatically on the base of the results of [27–34].

Generally speaking, instead of obtaining the solution to the Cauchy problem on a short time interval by some methods that are approximate anyway, which does not provide actual information of the real basic properties of the system, we determine the time interval, on which the Cauchy problem for the forward Kolmogorov system must really be solved and find this solution.

It is worth noting that exact estimates of the rate of convergence yield exact estimates of stability (perturbation bounds), see [8, 11, 14, 15, 17, 23, 32] and references therein. Moreover, such connections and their significance were highlighted in the recent communication by Mitrophanov, see

http://alexmitr.com/talk_DDE2018_Mitrophanov_FIN_post_sm.pdf.

The approach is based on the special properties of linear systems of differential equations with essentially nonnegative matrices. Specifically, if the column-wise sums of the elements of this matrix are identical and equal to, say, $-\alpha^*(t)$, then the exact upper bound of order $\exp\big\{-\int_0^t \alpha^*(u)\,du\big\}$ can be obtained for the rate of convergence of the solutions of the system in the corresponding metric. Moreover, if the column-wise sums of the absolute values of the elements of this matrix are identical and equal to, say, $\chi^*(t)$, then the exact lower bound of order $\exp\big\{-\int_0^t \chi^*(u)\,du\big\}$ can be obtained for the convergence rate as well. The bounds are obtained in three steps. At first step one excludes the (0) state from the forward Kolmogorov system of differential equations and thus obtains the new system with the new intensity matrix which is, in general, not non-diagonally non-negative. The second step is to transform the new intensity matrix in such a way that non-diagonally elements are non-negative and which leads to (loosely speaking) least distance between specifically defined upper and lower bounds. At third step one uses the logarithmic norm for the estimation of the convergence rate.

Here the key step is the second one. The transformation is made using a sequence of positive numbers $\{d_i, i \ge 1\}$, which does not have any probabilistic meaning and can be considered as an analogue of Lyapunov functions.

The advantages of this three-step approach is that it allows one to deal with time-homogeneous and time-inhomogeneous processes and it leads to exact both upper and lower bounds for the convergence rate. In time-homogeneous case the approach allows one to obtain the corresponding bounds for the decay parameter and gives an explicit bounds in total variation norm.

2 General Transformations

Recall that one has introduced $A(t)$ as the transposed intensity matrix $Q(t)$. Thus it has the form

$$A(t) = \begin{pmatrix} a_{00}(t) & a_{01}(t) & \cdots & a_{0r}(t) & \cdots \\ a_{10}(t) & a_{11}(t) & \cdots & a_{1r}(t) & \cdots \\ a_{20}(t) & a_{21}(t) & \cdots & a_{2r}(t) & \cdots \\ \cdots & & & & \\ a_{r0}(t) & a_{r1}(t) & \cdots & a_{rr}(t) & \cdots \\ \cdots & & & & \end{pmatrix}, \tag{4}$$

where $a_{ii}(t) = -\sum_{k\in\mathcal{X}, k\neq i} a_{ki}(t)$. Since $p_0(t) = 1 - \sum_{i=1}^{\infty} p_i(t)$ due to normalization condition, one can rewrite the system (3) as follows:

$$\frac{d}{dt}\mathbf{z}(t) = B(t)\mathbf{z}(t) + \mathbf{f}(t), \tag{5}$$

where

$$\mathbf{f}(t) = (a_{10}(t), a_{20}(t), \ldots)^T\ ,\ \mathbf{z}(t) = (p_1(t), p_2(t), \ldots)^T\ ,$$

$$B(t) = \begin{pmatrix} a_{11}-a_{10} & a_{12}-a_{10} & \cdots & a_{1r}-a_{10} & \cdots \\ a_{21}-a_{20} & a_{22}-a_{20} & \cdots & a_{2r}-a_{20} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{r1}-a_{r0} & a_{r2}-a_{r0} & \cdots & a_{rr}-a_{r0} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{6}$$

Each entry of B depends on t. See detailed discussion of this transformation in [7, 27].

There is the following simple relationship between pairs, $\mathbf{z}^{(i)} = \mathbf{z}^{(i)}(t),\ t \geq 0,\ i = 1, 2$, of solutions of (5) and pairs of solutions of (3), $\mathbf{p}^{(i)} = \mathbf{p}^{(i)}(t),\ t \geq 0,\ i = 1, 2$:

$$\begin{aligned} \left\|\mathbf{p}^{(1)} - \mathbf{p}^{(2)}\right\|_1 &= \left|p_0^{(1)} - p_0^{(2)}\right| + \sum_{i\geq 1}\left|p_i^{(1)} - p_i^{(2)}\right| \\ &= \left|1 - \sum_{i\geq 1} p_i^{(1)} - \left(1 - \sum_{i\geq 1} p_i^{(2)}\right)\right| + \left\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\right\|_1 \\ &= \left|\sum_{i\geq 1}\left(p_i^{(2)} - p_i^{(1)}\right)\right| + \left\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\right\|_1 \leq \sum_{i\geq 1}\left|p_i^{(2)} - p_i^{(1)}\right| + \left\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\right\|_1 \\ &= 2\left\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\right\|_1, \quad t \geq 0. \end{aligned}$$

Consequently,

$$\left\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\right\|_1 \leq \left\|\mathbf{p}^{(1)} - \mathbf{p}^{(2)}\right\|_1 \leq 2\left\|\mathbf{z}^{(1)} - \mathbf{z}^{(2)}\right\|_1, \quad t \geq 0, \tag{7}$$

which will be used in the study of stability and ergodicity.

Let $\{d_i,\ i \geq 1\}$ with $d_1 = 1$ be an increasing sequence of positive numbers. Put

$$W = \inf_{i\geq 1}\frac{d_i}{i}. \tag{8}$$

and denote by D the upper triangular matrix of the following form:

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots \\ 0 & d_2 & d_2 & \cdots \\ 0 & 0 & d_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{9}$$

Let l_{1D} be the corresponding space of sequences

$$l_{1D} = \left\{ \mathbf{z}(t) = (p_1(t), p_2(t), \cdots)^T \mid \|\mathbf{z}(t)\|_{1D} \equiv \|D\mathbf{z}(t)\|_1 < \infty \right\}$$

and introduce also the auxiliary norm $\|\cdot\|_{1E}$ defined as $\|\mathbf{z}(t)\|_{1E} = \sum_{k=1}^{\infty} k|p_k(t)|$. Then in $\|\cdot\|_{1D}$ norm the following two inequalities hold:

$$\begin{aligned}
\|\mathbf{z}(t)\|_{1D} = d_1 \left| \sum_{i=1}^{\infty} p_i(t) \right| + d_2 \left| \sum_{i=2}^{\infty} p_i(t) \right| \\
+ d_3 \left| \sum_{i=3}^{\infty} p_i(t) \right| + \dots \\
\geq \left(\left| \sum_{i=1}^{\infty} p_i(t) \right| + \left| \sum_{i=2}^{\infty} p_i(t) \right| + \left| \sum_{i=3}^{\infty} p_i(t) \right| + \dots \right) \\
\geq \frac{1}{2} \left(\left(\left| \sum_{i=1}^{\infty} p_i(t) \right| + \left| \sum_{i=2}^{\infty} p_i(t) \right| \right) \right. \\
\left. + \left(\left| \sum_{i=2}^{\infty} p_i(t) \right| + \left| \sum_{i=3}^{\infty} p_i(t) \right| \right) + \dots \right. \\
\geq \frac{1}{2} \sum_{i=1}^{\infty} |p_i(t)| = \frac{1}{2} \|\mathbf{z}(t)\|_1,
\end{aligned} \tag{10}$$

$$\begin{aligned}
\|\mathbf{z}(t)\|_{1E} = \sum_{k=1}^{\infty} k |p_k(t)| \\
= \sum_{k=1}^{\infty} \frac{k}{d_k} d_k |p_k(t)| \leq W^{-1} \sum_{k=1}^{\infty} d_k |p_k(t)| \\
= W^{-1} \sum_{k=1}^{\infty} d_k \left| \sum_{i=k}^{\infty} p_i(t) - \sum_{i=k-1}^{\infty} p_i(t) \right| \\
\leq W^{-1} \sum_{k=1}^{\infty} d_k \left(\left| \sum_{i=k}^{\infty} p_i(t) \right| + \left| \sum_{i=k-1}^{\infty} p_i(t) \right| \right) \\
\leq \frac{2}{W} \sum_{k=1}^{\infty} d_k \left| \sum_{i=k}^{\infty} p_i(t) \right| \leq \frac{2}{W} \|\mathbf{z}(t)\|_{1D}.
\end{aligned} \tag{11}$$

3 Logarithmic Norm and Related Bounds

Recall here the definition of logarithmic norm.

The concept of *logarithmic norm* of a square matrix was developed independently by Dahlquist [1] and Lozinskiĭ [12] as a tool to derive error bounds in the numerical integration of initial-value problems for a system of ordinary differential equations (see also the survey papers [21] and [20]). For the linear differential equation in a Banach space with locally integrable operator function this notion was discussed in [2].

Let $B(t)$, $t \geq 0$ be a one-parameter family of bounded linear operators on a Banach space $\mathscr{B}$ and let I denote the identity operator.
For each $t \geq 0$, the number

$$\gamma(B(t)) = \lim_{h \to +0} \frac{\|I + hB(t)\| - 1}{h} \tag{12}$$

is called the logarithmic norm of the operator $B(t)$.
The logarithmic norm of the matrix $B(t) = \left\{b_{ij}(t)\right\}$, $t \geq 0$ corresponding to a linear operator on the vector space $\mathscr{B}$ equipped with ℓ_1- norm, is

$$\gamma(B(t)) = \sup_j \left(b_{jj}(t) + \sum_{i \neq j} \left|b_{ij}(t)\right| \right), \quad t \geq 0. \tag{13}$$

Associate now the family of operators $B(t)$, $t \geq 0$ with the system of differential equations

$$\frac{d\mathbf{x}}{dt} = B(t)\mathbf{x}, \quad t \geq 0, \tag{14}$$

where the functions $b_{ij}(t)$, $0 \leq i, j < \infty$ are assumed to be locally integrable on $[0, \infty)$, and denote by $V(t,s)$, $0 \leq s \leq t$ the corresponding Cauchy operator (hence $\mathbf{x}(t) = V(t,s)\mathbf{x}(s)$ for any $0 \leq s \leq t$). Then the logarithmic norm of the operator $B(t)$ is related to $V(t,s)$, $0 \leq s \leq t$ by

$$\gamma(B(t)) = \lim_{h \to +0} \frac{\|V(t+h,t)\| - 1}{h}, \quad t \geq 0. \tag{15}$$

From the latter one can deduce the following bounds on the $\mathscr{B}$-norm of the Cauchy operator $V(t,s)$, $0 \leq s \leq t$:

$$e^{-\int_s^t \gamma(-B(\tau))\, d\tau} \leq \|V(t,s)\| \leq e^{\int_s^t \gamma(B(\tau))\, d\tau}, \quad 0 \leq s \leq t. \tag{16}$$

Moreover, for any solution $\mathbf{x}(t) \in \mathscr{B}$, $t \geq 0$ of (14) we have

$$\|\mathbf{x}(t)\| \geq e^{-\int_s^t \gamma(-B(\tau))\, d\tau} \|\mathbf{x}(s)\|. \tag{17}$$

We will also make use of the fact that if $\mathscr{B}$ is a vector space with norm ℓ_1 and all diagonal elements of B are non-negative then, by (13)

$$\gamma\left(B\left(t\right)\right) = \sup_j \sum_i b_{ij}\left(t\right), \quad t \geq 0,$$

and, *a fortiori*, for any solution $\mathbf{x}\left(t\right),\ t \geq 0$ of (14), s.t. $\mathbf{x}\left(s\right) \geq \mathbf{0}$, we have

$$\left\|\mathbf{x}\left(t\right)\right\| \geq e^{\int\limits_s^t \inf\limits_j \sum\limits_i b_{ij}(\tau)\ d\tau} \left\|\mathbf{x}\left(s\right)\right\|, \quad 0 \leq s \leq t. \tag{18}$$

Consider the Eq. (5) in the space l_{1D}, where $B(t)$ and $\mathbf{f}(t)$ are locally integrable on $[0, +\infty)$. Let one compute the logarithmic norm of operator function $B(t)$.

Then for the logarithmic norm of the operator function $B(t)$ in $\|\cdot\|_{1D}$ norm the following equality holds:

$$\gamma(B(t))_{1D} = \gamma(DB(t)D^{-1})_1.$$

Denote by $B^*(t) = DB(t)D^{-1}$, and the elements of $B^*(t)$ by $b^*_{ij}(t)$ i.e. $B^*(t) = \left(b^*_{ij}(t)\right)_{i,j=1}^{\infty}$. Assume that

$$b^*_{ij}(t) \geq 0,\ i \neq j,\ t \geq 0. \tag{19}$$

Remark 1. Note that assumption (19) of essential nonnegativity of the reduced matrix $B^*(t)$ is key to the possibility of effective use of the method of the logarithmic norm. In particular, this assumption is fulfilled for four important classes of Markov chains, which we consider in the next section.

Put

$$\alpha_i\left(t\right) = -\sum_{j=0}^{\infty} b^*_{ji}(t), \quad \chi_i\left(t\right) = -\sum_{j=0}^{\infty} |b^*_{ji}(t)|,\ i \geq 1, \tag{20}$$

and let $\alpha(t)$ and $\beta(t)$ denote the least lower and the least upper bound of the sequence of functions $\{\alpha_i(t),\ i \geq 1\}$ and $\chi(t)$ denote the least upper bound of $\{\chi_i(t),\ i \geq 1\}$ i.e.

$$\alpha\left(t\right) = \inf_{i \geq 1} \alpha_i\left(t\right),\ \beta\left(t\right) = \sup_{i \geq 1} \alpha_i\left(t\right), \tag{21}$$

$$\chi\left(t\right) = \sup_{i \geq 1} \chi_i\left(t\right). \tag{22}$$

Then the logarithmic norms of $B(t)$ and $(-B(t))$ are equal to

$$\gamma\left(B\left(t\right)\right)_{1D} = \sup_i \alpha_i(t) = -\alpha\left(t\right),$$
$$\gamma\left(-B\left(t\right)\right)_{1D} = \sup \chi_i\left(t\right) = \chi\left(t\right).$$

If now one defines $\mathbf{v}(t) = D\left(\mathbf{p}^{*}(t) - \mathbf{p}^{**}(t)\right)$, then the following equation holds

$$\frac{d}{dt}\mathbf{v}(t) = DB(t)\,D^{-1}\mathbf{v}(t), \tag{23}$$

Notice that due to (19), the inequality $\mathbf{v}(s) \geq \mathbf{0}$ implies that $\mathbf{v}(t) \geq \mathbf{0}$ for any $t \geq s$. Hence

$$\frac{d}{dt}\sum_{i=1}^{\infty} v_i(t) \geq -\beta(t)\sum_{i=1}^{\infty} v_i(t), \tag{24}$$

and one can obtain establish the corresponding bounds on the rate of convergence, perturbation bounds, and estimates on the error of truncations.

4 Four Classes of Markov Chains

These classes were previously studied in [34, 35]. We use here the terminology from Markov chain theory and queueing in parallel depending on context.

Class (I). Inhomogeneous birth-death processes (BDP), where all $a_{ij}(t) = 0$ for any $t \geq 0$ if $|i-j| > 1$, and $a_{i,i+1}(t) = \mu_{i+1}(t),\ \ a_{i+1,i}(t) = \lambda_i(t)$ - birth and death rates respectively. This process, in particular, is a standard model as queue-length process for a general Markovian queue $M_n(t)/M_n(t)/1$.

In this situation we obtain

$$B^{*}(t) = \begin{pmatrix} -\left(\lambda_0(t)+\mu_1(t)\right) & \mu_1(t) & 0 & \cdots & 0 & \cdots & \cdots \\ \lambda_1(t) & -\left(\lambda_1(t)+\mu_2(t)\right) & \mu_2(t) & \cdots & 0 & \cdots & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \\ 0 & \cdots & \cdots & \lambda_{r-1}(t) & -\left(\lambda_{r-1}(t)+\mu_r(t)\right) & \mu_r(t) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \tag{25}$$

if $S = \infty$, and

$$B^{*}(t) = \begin{pmatrix} -\left(\lambda_0(t)+\mu_1(t)\right) & \mu_1(t) & 0 & \cdots & 0 \\ \lambda_1(t) & -\left(\lambda_1(t)+\mu_2(t)\right) & \mu_2(t) & \cdots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & \lambda_{S-1}(t) & -\left(\lambda_{S-1}(t)+\mu_S(t)\right) \end{pmatrix}, \tag{26}$$

if $S < \infty$.

One can see that the transformed matrix $B^{*}(t)$ is essentially nonnegative for any t, that is all off-diagonal elements of this matrix are nonnegative for any t.

Remark 2. This class is the most studied. It includes, in particular, models of systems of the theory of queues $M_t/M_t/N$, and $M_t/M_t/N/N$, see for instance [3–5, 7, 13, 22, 25–27, 30] and references therein. For the first one, we get the matrix (25) with

$\lambda_k(t) = \lambda(t)$ and $\mu_k(t) = \min(k, N) \cdot \mu(t)$, and for the second one we get (26) with $\lambda_k(t) = \lambda(t)$ and $\mu_k(t) = k\mu(t)$.

Another approach to the study of close models with discrete time was considered in [9].

Class (II). Inhomogeneous queue-length process for a queue with batch arrivals and single services, where $a_{ij}(t) = 0$ for any $t \geq 0$ if $i < j - 1$, all arrival rates do not depend on the size of a queue, where $a_{i+k,i}(t) = a_k(t)$ for $k \geq 1$ - the rate of arrival of a group of k customers, $a_{i,i+1}(t) = \mu_{i+1}(t)$ - the service rate. Such models in simplest situations were firstly considered in [16].

In this situation we have

$$B^*(t) = \begin{pmatrix} a_{11}(t) & \mu_1(t) & 0 & \cdots & 0 \\ a_1(t) & a_{22}(t) & \mu_2(t) & \cdots & 0 \\ a_2(t) & a_1(t) & a_{33}(t) & \mu_3(t) & \cdots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{27}$$

if $S = \infty$, and

$$B^*(t) = \begin{pmatrix} a_{11}(t) - a_S(t) & \mu_1(t) & 0 & \cdots & 0 \\ a_1(t) - a_S(t) & a_{22}(t) - a_{S-1}(t) & \mu_2(t) & \cdots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ a_{S-1}(t) - a_S(t) & \cdots & \cdots & a_1(t) - a_2(t) & a_{SS}(t) - a_1(t) \end{pmatrix}, \tag{28}$$

if $S < \infty$.

One can see that the transformed matrix $B^*(t)$ is certainly essentially nonnegative for any t if arrival rates $a_k(t)$ are decrease in k.

Class (III). Inhomogeneous queue-length process for the queueing model with batch services and single arrivals, where all $a_{ij}(t) = 0$ for any $t \geq 0$ if $i > j + 1$, and all service rates do not depend on the size of a queue, where $a_{i,i+k}(t) = b_k(t),\ k \geq 1$ is the rate of service of a group of k customers, and $a_{i+1,i}(t) = \lambda_i(t)$ is the arrival rate, see also [16]. One can find more modern studies of these models in [10].

Here we obtain

$$B^*(t) = \begin{pmatrix} -\left(\lambda_0(t) + b_1(t)\right) & b_1(t) - b_2(t) & b_2(t) - b_3(t) & \cdots & \cdots & \\ \lambda_1(t) & -\bigl(\lambda_1(t) + \sum\limits_{i \leq 2} b_i(t)\bigr) & b_1(t) - b_3(t) & \cdots & \cdots & \\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & \cdots & \cdots & \lambda_{r-1}(t) & -\bigl(\lambda_{r-1}(t) + \sum\limits_{i \leq r} b_i(t)\bigr) \cdots & \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{29}$$

if $S = \infty$, and

$$B^*(t) = \begin{pmatrix} -(\lambda_0(t) + b_1(t)) & b_1(t) - b_2(t) & b_2(t) - b_3(t) & \cdots & b_{S-1}(t) - b_S(t) \\ \lambda_1(t) & -\big(\lambda_1(t) + \sum\limits_{i \le 2} b_i(t)\big) & b_1(t) - b_3(t) & \cdots & b_{S-2}(t) - b_S(t) \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & \lambda_{S-1}(t) & -\big(\lambda_{S-1}(t) + \sum\limits_{i \le S} b_i(t)\big) \end{pmatrix}, \quad (30)$$

if $S < \infty$.

One can see that the transformed matrix $B^*(t)$ is certainly essentially nonnegative for any t if service rates $b_k(t)$ are decrease in k.

Class (IY). Queue-length process for a non-stationary queueing model with batch arrivals and group services, where all rates do not depend on the size of a queue, here $a_{i+k,i}(t) = a_k(t)$, and $a_{i,i+k}(t) = b_k(t)$ for $k \ge 1$ are the rates of arrival and service of a group of k customers respectively. Such process were studied in [18, 19, 31].

$$B^* = \begin{pmatrix} a_{11}(t) & b_1(t) - b_2(t) & b_2(t) - b_3(t) & \cdots & \cdots & \\ a_1(t) & a_{22}(t) & b_1(t) - b_3(t) & \cdots & \cdots & \\ & & & & & \\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ a_{r-1}(t) & \cdots & \cdots & a_1(t) & a_{rr}(t) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \quad (31)$$

if $S = \infty$, and

$$B^*(t) = \begin{pmatrix} a_{11}(t) - a_S(t) & b_1(t) - b_2(t) & b_2(t) - b_3(t) & \cdots & b_{S-1}(t) - b_S(t) \\ a_1(t) - a_S(t) & a_{22}(t) - a_{S-1}(t) & b_1(t) - b_3(t) & \cdots & b_{S-2}(t) - b_S(t) \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ a_{S-1}(t) - a_S(t) & \cdots & \cdots & a_1(t) - a_2(t) & a_{SS}(t) - a_1(t) \end{pmatrix}, \quad (32)$$

if $S < \infty$.

In this case the transformed matrix $B^*(t)$ is surely essentially nonnegative for any t if all arrival and service rates $a_k(t)$ and $b_k(t)$ are decreasing on k.

5 General Bounds for Continuous-Time Markov Chains

Rate of Convergence

Theorem 1. *Let there exist an increasing sequence $\{d_j,\ j \ge 1\}$ of positive numbers with $d_1 = 1$, such that (19) holds, and $\alpha(t)$ defined by (21) satisfies*

$$\int_0^\infty \alpha(t)\,dt = +\infty. \tag{33}$$

Then the Markov chain $\{X(t),\ t \geq 0\}$ *is weakly ergodic and the following bounds hold:*

$$\begin{aligned} & e^{-\int_s^t \chi(u)du} \|\mathbf{p}^*(s) - \mathbf{p}^{**}(s)\|_{1D} \\ & \quad \leq \|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} \\ & \leq e^{-\int_s^t \alpha(u)du} \|\mathbf{p}^*(s) - \mathbf{p}^{**}(s)\|_{1D}, \end{aligned} \tag{34}$$

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \leq 4e^{-\int_s^t \alpha(u)du} \|\mathbf{z}^*(s) - \mathbf{z}^{**}(s)\|_{1D}, \tag{35}$$

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1E} \leq \frac{2}{W} e^{-\int_s^t \alpha(u)du} \|\mathbf{z}^*(s) - \mathbf{z}^{**}(s)\|_{1D}, \tag{36}$$

for any initial conditions $s \geq 0$, $\mathbf{p}^*(s)$, $\mathbf{p}^{**}(s)$ *and any* $t \geq s$.

If in addition $D\left(\mathbf{p}^*(s) - \mathbf{p}^{**}(s)\right) \geq \mathbf{0}$, *then*

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} \geq e^{-\int_s^t \beta(u)du} \|\mathbf{p}^*(s) - \mathbf{p}^{**}(s)\|_{1D}, \tag{37}$$

for any $0 \leq s \leq t$.

If the Markov chain is homogeneous, then all elements $b^*_{ij}(t)$ of the matrix $DB(t)D^{-1}$ do not dependent on t i.e. the quantities in (21) are constants. Thus instead of general bounds given by Theorem 1, one can specify then and obtain the following theorem.

Theorem 2. *Let there exist an increasing sequence* $\{d_j,\ j \geq 1\}$ *of positive numbers with* $d_1 = 1$, *such that (19) holds, and* $\alpha(t) = \alpha$ *defined by (21) is positive i.e.* $\alpha > 0$. *Then the Markov chain* $\{X(t),\ t \geq 0\}$ *is strongly ergodic and the following bounds hold:*

$$\begin{aligned} e^{-\chi t} \|\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\|_{1D} & \leq \|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} \\ & \leq e^{-\alpha t} \|\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\|_{1D}, \end{aligned} \tag{38}$$

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \leq 4e^{-\alpha t} \|\mathbf{z}^*(0) - \mathbf{z}^{**}(0)\|_{1D}, \tag{39}$$

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1E} \leq \frac{2}{W} e^{-\alpha t} \|\mathbf{z}^*(0) - \mathbf{z}^{**}(0)\|_{1D}, \tag{40}$$

for any initial conditions $s \geq 0$, $\mathbf{p}^*(0)$, $\mathbf{p}^{**}(0)$ *and any* $t \geq 0$.

If in addition $D\left(\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\right) \geq \mathbf{0}$, *then*

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\|_{1D} \geq e^{-\beta t} \|\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\|_{1D}, \tag{41}$$

for any $t \geq 0$.

For the decay parameter α^* *defined as*

$$\lim_{t\to\infty}(p_{ij}(t)-\pi_j)=O(e^{-\alpha^* t}),$$

where $\{\pi_j,\ j\geq 0\}$ *are the stationary probabilities of the chain, it holds that* $\alpha^*\geq\alpha$.

Notice that some additional results related to *Theorem* 2 can also be found in [4, 6]. If one assumes that the intensities $q_{ij}(t)$ are $1-$periodic in t i.e. $q_{ij}(t)$ are periodic functions and the length of the period is equal to one, then the Markov chain $\{X(t),\ t\geq 0\}$ has the limiting $1-$periodic limiting regime. Under the assumptions of Theorem 1 the Markov chain $\{X(t),\ t\geq 0\}$ is exponentially weakly ergodic. The detailed discussion of this results is given in [27].

Consider now a bit more detailed analysis of two special cases: homogeneous case and the case with periodic intensities. Firstly note that in the both cases there exist positive M and a such that

$$e^{-\int_s^t\alpha(u)\,du}\leq Me^{-a(t-s)} \tag{42}$$

for any $0\leq s\leq t$. Hence the Markov chain $\{X(t),\ t\geq 0\}$ is *exponentially* weakly ergodic. Indeed, if the Markov chain $\{X(t),\ t\geq 0\}$ is homogeneous, then one may put $M=1$, $a=\alpha$ given by (21). If all the intensity functions $q_{ij}(t)$ are $1-$periodic in t, then one may put

$$a=\int_0^1\alpha(t)\,dt,\quad M=e^K,\quad K=\sup_{|t-s|\leq 1}\int_s^t\alpha(u)\,du.$$

By doing so, for any solution of (5) the following bound holds:

$$\begin{aligned}&\|\mathbf{z}(t)\|_{1D}\\ \leq\ \|V(t)\|_{1D}\|\mathbf{z}(0)\|_{1D}+\int_0^t\ &\|V(t,\tau)\|_{1D}\|\mathbf{f}(\tau)\|_{1D}\,d\tau\\ &\leq\ Me^{-at}\|\mathbf{z}(0)\|_{1D}+\frac{FM}{a},\end{aligned} \tag{43}$$

where F is such that $\|\mathbf{f}(t)\|_{1D}\leq F$ for almost all $t\in[0,1]$. Hence one has the upper bound for the limit

$$\limsup_{t\to\infty}\|\mathbf{z}(t)\|_{1D}\leq\frac{FM}{a}, \tag{44}$$

for any initial condition and

$$\|\mathbf{p}(0)-\mathbf{e}_0\|_{1D}=\|\mathbf{p}(0)\|_{1D}=\|\mathbf{z}(0)\|_{1D}\leq\limsup_{t\to\infty}\|\mathbf{z}(t)\|_{1D}, \tag{45}$$

where $\mathbf{e}_i$ denotes the unit vector of zeros with 1 in the i-th place. If the initial distribution is $\mathbf{p}^{**}(0) = \mathbf{e}_0$ then $\mathbf{z}^{**}(0) = \mathbf{0}$, $\mathbf{z}(t) \geq 0$ for any $\mathbf{p}^*(0)$ and any $t \geq 0$. Therefore

$$\begin{aligned} \|\mathbf{z}(t)\|_{1D} &= d_1 p_1 + (d_1 + d_2) p_2 \\ &+ (d_1 + d_2 + d_3) p_3 + ... \\ = d_1 p_1 + \frac{d_1 + d_2}{2} 2 p_2 + \frac{d_1 + d_2 + d_3}{3} 3 p_3 + ... \\ &\geq \inf_k \frac{d_1 + ... + d_k}{k} \|\mathbf{z}(t)\|_{1E}, \end{aligned}$$

and one can use $W^* = \inf_k \frac{d_1 + ... + d_k}{k}$ instead of $W = \inf_k \frac{d_k}{k}$, given by (8) in all the bounds on the rate of convergence. Finally, for the considered two special cases one has the following two corollaries.

Corollary 1. *Let $\{X(t),\ t \geq 0\}$ be a homogeneous Markov chain and let there exist an increasing sequence $\{d_j,\ j \geq 1\}$ of positive numbers with $d_1 = 1$ such that (19) holds and in addition $\alpha > 0$. Then the Markov chain $\{X(t),\ t \geq 0\}$ is exponentially ergodic and the following bounds hold:*

$$\|\pi - \mathbf{p}(t, 0)\| \leq \frac{4F}{\alpha} e^{-\alpha t}, \tag{46}$$

$$|\varphi - E(t, 0)| \leq \frac{F}{\alpha W^*} e^{-\alpha t}, \tag{47}$$

where $\pi = (\pi_0, \pi_1, \dots)^T$ denotes the vector of stationary probabilities of the chain and $\varphi = \sum_{j=0}^{\infty} j \pi_j$ and $\mathbf{p}(0, 0) = \mathbf{e}_0$.

Corollary 2. *Assume that all the intensity functions of the Markov chain $\{X(t),\ t \geq 0\}$ are $1-$periodic in t. Let there exist an increasing sequence $\{d_j,\ j \geq 1\}$ of positive numbers with $d_1 = 1$ such that (19) holds and in addition $\int_0^1 \alpha(t)\, dt = a > 0$. Then the Markov chain $\{X(t),\ t \geq 0\}$ is exponentially weakly ergodic and the following bounds hold:*

$$\|\pi(t) - \mathbf{p}(t, 0)\| \leq \frac{4FM}{a} e^{-at}, \tag{48}$$

$$|\varphi(t) - E(t, 0)| \leq \frac{FM}{aW^*} e^{-at}, \tag{49}$$

where $\pi(t) = (\pi_0(t), \pi_1(t), \dots)^T$ denotes the vector of limiting probabilities of the chain and $\varphi(t) = \sum_{j=0}^{\infty} j \pi_j(t)$ and $\mathbf{p}(0, 0) = \mathbf{e}_0$.

If the state space of the Markov chain is finite there exist a number of special results (see [4, 6, 29]).

Perturbation Bounds

Let $\{\bar{X}(t), t \geq 0\}$ be a perturbed Markov chain with transposed intensity matrix $\bar{A}(t)$ and the same restrictions as imposed on $A(t)$. It is assumed that $\|\hat{A}(t)\| = \|A(t) - \bar{A}(t)\| \leq \varepsilon$, for almost all $t \geq 0$, which means the perturbations are considered to be small in l_1 norm.

We can obtain the corresponding perturbation bounds. There are two different approaches.

The first approach in this direction are given in [8, 23] both for the discrete and continuous time Markov chains respectively. In the considered situation of Markov chains with continuous time, this approach is based on a comparison of the Cauchy operators of two linear equations in a Banach space considered in [2]. Consider Eq. (5) for the perturbed chain:

$$\frac{d}{dt}\bar{\mathbf{z}}(t) = \bar{B}(t)\bar{\mathbf{z}}(t) + \bar{\mathbf{f}}(t). \tag{50}$$

In this case, the weight space l_{1D} is considered as the base one, and the norms of perturbations are assumed to be small both in l_1 and l_{1D} norms. Namely, we suppose that $\|\hat{B}(t)\|_{1D} = \|B(t) - \bar{B}(t)\|_{1D} \leq \varepsilon$, and $\|\mathbf{f}(t) - \bar{\mathbf{f}}(t)\|_{1D} \leq \varepsilon$, for almost all $t \geq 0$.

The corresponding general results have been obtained in [32]. A typical statement of this kind is as follows:

Theorem 3. *Let the assumptions of Theorem 1 be fulfilled, and let, an addition, $X(t)$ be exponentially weakly ergodic in l_{1D} norm with the corresponding parameters M_D, a_D in (42). Then the following perturbation bound holds:*

$$\limsup_{t\to\infty} \|\mathbf{p}(t) - \bar{\mathbf{p}}(t)\|_1 \leq \frac{4M_D\varepsilon\,(M_D\mathsf{F} + a_D)}{a_D\,(a_D - M_D\varepsilon)}, \tag{51}$$

where $\|\bar{\mathbf{f}}(t)\|_{1D} \leq \mathsf{F}$ for almost all $t \geq 0$.

The second approach also began with [23], namely, Mitrophanov [14] successfully applied probabilistic considerations and ergodicity in uniform operator topology which allowed to significantly reduce the constant factor in the stability estimate. The corresponding bounds for inhomogeneous situation has been obtained in [28].

A typical statement of this kind is as follows:

Theorem 4. *Let Markov cain $X(t)$ be exponentially weakly ergodic in l_1 norm with the corresponding parameters M^*, α^* in (42). Then the following bound holds:*

$$\limsup_{t\to\infty} \|\mathbf{p}(t) - \bar{\mathbf{p}}(t)\|_1 \leq \frac{\varepsilon\,(1 + \log M^*)}{\alpha^*}. \tag{52}$$

Truncation Bounds

Calculation of the limiting characteristics for (inhomogeneous) birth-death processes via truncations was firstly mentioned in [24] and was considered in details in [27]. First results for more general Markovian queueing models have been obtained recently in [31]. The respective bound of approximation error as a rule depends on time. Vladimir V. Kalashnikov in the early 1990-s suggested that in some cases one can obtain uniform (in time) error bounds of truncation. Such bounds for inhomogeneous birth-death processes have been obtained in [30], and for more general Markov chains in [33], this statement can be formulated in the following way.

Let $X_{N-1}(t)$ be a truncated process with the state space $\mathrm{E}_{N-1} = \{0, 1, \dots, N-1\}$ and the corresponding transposed infinitesimal matrix

$$A_{N-1}(t) = \begin{pmatrix} b_{00}(t) & a_{01}(t) & \cdots & a_{0,N-1}(t) \\ a_{10}(t) & b_{11}(t) & \cdots & a_{1,N-1}(t) \\ a_{20}(t) & a_{21}(t) & \cdots & a_{2,N-1}(t) \\ \multicolumn{4}{c}{\dotfill} \\ a_{N-1,0}(t) & a_{N-1,1}(t) & \cdots & b_{N-1,N-1}(t) \end{pmatrix},$$

where $b_{ii}(t) = -\sum_{k=0,k\neq i}^{N-1} a_{ki}(t)$.

Then the forward Kolmogorov system for $X_{N-1}(t)$ is

$$\frac{d\mathbf{p}^*}{dt} = A_{N-1}(t)\mathbf{p}^*,$$

and instead of (5) we have

$$\frac{d\mathbf{z}^*}{dt} = B_{N-1}(t)\mathbf{z}^*(t) + \mathbf{f}_{N-1}(t), \tag{53}$$

where $\mathbf{f}_{N-1}(t) = (a_{10}(t), a_{20}(t), \dots, a_{N-1,0}(t))^\top$, $\mathbf{z}^*(t) = (p_1, p_2, \cdots, p_{N-1})^\top$,

$$B_{N-1} = \begin{pmatrix} b_{11}(t) - a_{10}(t) & a_{12}(t) - a_{10}(t) & \cdots & a_{1,N-1}(t) - a_{10}(t) \\ a_{21}(t) - a_{20}(t) & b_{22}(t) - a_{20}(t) & \cdots & a_{2,N-1}(t) - a_{20}(t) \\ \multicolumn{4}{c}{\dotfill} \\ a_{N-1,1}(t) - a_{N-1,0}(t) & a_{N-1,2}(t) - a_{N-1,0}(t) & \cdots & b_{N-1,N-1}(t) - a_{N-1,0}(t) \end{pmatrix}.$$

Below we will identify the finite vector with entries $(a_1, \dots, a_{N-1})^\top$ and the infinite vector with the same first $N-1$ coordinates and the others equal to zero. Moreover we suppose that

$$a_{i+k,i}(t) = q_{i,i+k}(t) \le R \cdot q^{-k}, \qquad q > 1, \quad R > 0, \tag{54}$$

for any $k \ge 1$, $i \ge 0$ and almost all $t \ge 0$. For $\delta \in (1, \sqrt{q})$ we consider the sequences $d_k = \delta^{k-1}$ and $d_k^* = \delta^{2k-2}$, $k \ge 1$.

Denote

$$W = \inf_{i \ge 1} \frac{d_i}{i}, \qquad g_i = \sum_{n=1}^{i} d_n.$$

Let D and D^* be upper triangular matrices:

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots \\ 0 & d_2 & d_2 & \cdots \\ 0 & 0 & d_3 & \cdots \\ & \ddots & \ddots & \ddots \end{pmatrix}, \quad D^* = \begin{pmatrix} d_1^* & d_1^* & d_1^* & \cdots \\ 0 & d_2^* & d_2^* & \cdots \\ 0 & 0 & d_3^* & \cdots \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

and l_{1D}, l_{1D^*} be the corresponding spaces of sequences:

$$\begin{aligned} l_{1D} &= \{\mathbf{z} = (p_1, p_2, \ldots)^\top \mid \|\mathbf{z}\|_{1D} \equiv \|D\mathbf{z}\|_1 < \infty\}, \\ l_{1D^*} &= \{\mathbf{z} = (p_1, p_2, \ldots)^\top \mid \|\mathbf{z}\|_{1D^*} \equiv \|D^*\mathbf{z}\|_1 < \infty\}. \end{aligned}$$

We suppose that there exist positive constants M, a, M^*, a^* such that the following bounds

$$\|V(t,s)\|_{1D} \le Me^{-a(t-s)}, \tag{55}$$

and

$$\|V(t,s)\|_{1D^*} \le M^* e^{-a^*(t-s)}, \tag{56}$$

hold for Cauchy operator $V(t,s)$ of Eq. (5) for any s, t $(0 \le s \le t)$. These estimates guarantee exponential convergence to zero as $t - s \to \infty$ of the difference $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \to 0$ in l_{1D} and l_{1D^*} norms respectively for the corresponding initial conditions.

Theorem 5. *Let the assumptions* (54), (55), (56) be fulfilled. *Then the following bounds of error of truncations hold:*

$$\|\mathbf{p}(t) - \mathbf{p}_{N-1}(t)\| \le C_1 \Big(\frac{\delta^2}{q}\Big)^{N/3} + C_2 \delta^{-N/3} + C_3 \Big(\frac{\delta}{q}\Big)^N \tag{57}$$

and

$$|E(t,0) - E_{N-1}(t,0)| \le \frac{1}{W}\Big\{C_1 \Big(\frac{\delta^2}{q}\Big)^{N/3} + C_2 \delta^{-N/3} + C_3 \Big(\frac{\delta}{q}\Big)^N\Big\}, \tag{58}$$

where index $N-1$ shows the corresponding characteristics of truncated process and $X(0) = X_{N-1}(0) = 0$. Moreover, constants $C_i = C_i(\delta, q)$ do not depend on N and on $\delta \in (1, \sqrt{q})$.

Finally we can briefly describe the possible procedure for finding $\pi(t)$ and $\varphi(t)$ in case of $1-$periodic in t intensities. Firstly we estimate the instant $t = t^*$ (using the ergodicity bounds), starting from which the solution of the forward Kolmogorov system (3) with the initial condition $X(0)$ is within the fixed $\epsilon > 0$ from the limiting periodic probabilities. Then we estimate the size n^* of the state space $\{0, 1, \dots, n^*\}$, which guarantees the desired approximation error on the interval $[0, t^* + 1]$. Then we find the solution of the truncated system on the interval $[0, t^* + 1]$, eventually the values for $\pi(t)$ and $\varphi(t)$ on the interval $[t^*, t^* + 1]$.

Acknowledgements The research has been supported by the Russian Science Foundation under grant 19-11-00020. The author also would like to thank the organizers of ICDDEA 2019, for their hospitality.

References

1. Dahlquist, G.: Stability and error bounds in the numerical integration of ordinary differential equations. Inaugural dissertation, University of Stockholm, Almqvist & Wiksells Boktryckeri AB, Uppsala 1958. Reprinted in: Transactions of the Royal Institute of Technology, **130**, Stockholm (1959)
2. Daleckij, J.L., Krein, M.G.: Stability of solutions of differential equations in Banach space. Am. Math. Soc. Transl. **43** (2002)
3. Van Doorn, E.A., Zeifman, A.I.: On the speed of convergence to stationarity of the Erlang loss system. Queueing Syst. **63**, 241–252 (2009)
4. Van Doorn, E.A., Zeifman, A.I., Panfilova, T.L.: Bounds and asymptotics for the rate of convergence of birth-death processes. Theory Probab. Appl. **54**, 97–113 (2010)
5. Fricker, C., Robert, P., Tibi, D.: On the rate of convergence of Erlang's model. J. Appl. Probab. **36**, 1167–1184 (1999)
6. Gran ovsky, B.L., Zeifman, A.I.: The N-limit of spectral gap of a class of birth-death Markov chains. Appl. Stochast. Models Bus. Ind. **16**(4), 235–248 (2000)
7. Granovsky, B.L., Zeifman, A.: Nonstationary queues: estimation of the rate of convergence. Queueing Syst. **46**(3–4), 363–388 (2004)
8. Kartashov, N.V.: Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. Theory Probab. Appl. **30**, 71–89 (1985)
9. Kloeden, P.E., Kozyakin, V.: Asymptotic behaviour of random tridiagonal Markov chains in biological applications. Discrete Conts. Dyn. Syst. Ser. B **18**, 453–465 (2012)
10. Li, J., Zhang, L.: M X/M/c queue with catastrophes and state-dependent control at idle time. Front. Math. China **12**(6), 1427–1439 (2017)
11. Liu, Y.: Perturbation bounds for the stationary distributions of Markov chains. SIAM J. Matrix Anal. Appl. **33**(4), 1057–1074 (2012)
12. Lozinskiĭ, S. M.: Error estimate for numerical integration of ordinary differential equations, I. Izv. Vysš. Učebn. Zaved. Matematika **5**, 52-90 (1958) . Errata, **5** 222 (1959). (In Russian)
13. Margolius, B.: Periodic solution to the time-inhomogeneous multi-server Poisson queue. Oper. Res. Lett. **35**(1), 125–138 (2007)
14. Mitrophanov, A.Y.: Stability and exponential convergence of continuous-time Markov chains. J. Appl. Probab. **40**, 970–979 (2003)
15. Mitrophanov, A.Y.: The spectral gap and perturbation bounds for reversible continuous-time Markov chains. J. Appl. Probab. **41**, 1219–1222 (2004)
16. Nelson, R., Towsley, D., Tantawi, A.N.: Performance analysis of parallel processing systems. IEEE Trans. Softw. Eng. **14**(4), 532–540 (1988)

17. Rudolf, D., Schweizer, N.: Perturbation theory for Markov chains via Wasserstein distance. Bernoulli **24**(4A), 2610–2639 (2018)
18. Satin, Y.A., Zeifman, A.I., Korotysheva, A.V., Shorgin, S.Y.: On a class of Markovian queues. Inform. Appl. **5**(4), 18–24 (2011). (in Russian)
19. Satin, Y.A., Zeifman, A.I., Korotysheva, A.V.: On the rate of convergence and truncations for a class of Markovian queueing systems. Theory Probab. Appl. **57**, 529–539 (2013)
20. Söderlind, G.: The logarithmic norm. History and modern theory. BIT. Numer. Math. **46**, 631–652 (2006)
21. Ström, T.: On logarithmic norms. SIAM J. Numer. Anal. **12**, 741–753 (1975)
22. Voit, M.: A note of the rate of convergence to equilibrium for Erlang's model in the subcritical case. J. Appl. Probab. **37**, 918–923 (2000)
23. Zeifman, A.I.: Stability for continuous-time nonhomogeneous Markov chains. In: Kalashnikov, V.V., Zolotarev, V.M. (eds.) Stability Problems for Stochastic Models, pp. 401–414. Springer, Heidelberg (1985)
24. Zeifman, A.I.: Truncation error in a birth and death system. USSR Comput. Math. Math. Phys. **28**(6), 210–211 (1988)
25. Zeifman, A.I.: Some properties of a system with losses in the case of variable rates. Autom. Remote Control. **50**(1), 82–87 (1989)
26. Zeifman, A.I.: Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. Stochast. Process. Appl. **59**, 157–173 (1995)
27. Zeifman, A., Leorato, S., Orsingher, E., Satin, Y., Shilova, G.: Some universal limits for nonhomogeneous birth and death processes. Queueing Syst. **52**(2), 139–151 (2006)
28. Zeifman, A.I., Korotysheva, A.: Perturbation bounds for $M_t/M_t/N$ queue with catastrophes. Stochastic Models **28**, 49–62 (2012)
29. Zeifman, A., Satin, Y., Panfilova, T.: Limiting characteristics for finite birth-death-catastrophe processes. Math. Biosci. **245**(1), 96–102 (2013)
30. Zeifman, A., Satin, Y., Korolev, V., Shorgin, S.: On truncations for weakly ergodic inhomogeneous birth and death processes. Int. J. Appl. Math. Comput. Sci. **24**, 503–518 (2014)
31. Zeifman, A., Korotysheva, A., Korolev, V., Satin, Y., Bening, V.: Perturbation bounds and truncations for a class of Markovian queues. Queueing Syst. **76**, 205–221 (2004)
32. Zeifman, A.I., Korolev, V.Y.: On perturbation bounds for continuous-time Markov chains. Stat. Probab. Lett. **88**, 66–72 (2014)
33. Zeifman, A.I., Korotysheva, A.V., Korolev, V.Y., Satin, Y.A.: Truncation bounds for approximations of inhomogeneous continuous-time Markov chains. Theory Probab. Appl. **61**, 513–520 (2017)
34. Zeifman, A., Razumchik, R., Satin, Y., Kiseleva, K., Korotysheva, A., Korolev, V.: Bounds on the rate of convergence for one class of inhomogeneous Markovian queueing models with possible batch arrivals and services. Int. J. Appl. Math. Comput Sci. **28**, 141–154 (2018)
35. Zeifman, A., Sipin, A., Korolev, V., Shilova, G., Kiseleva, K., Korotysheva, A., Satin, Y.: On sharp bounds on the rate of convergence for finite continuous-time markovian queueing models. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A. (eds.) Computer Aided Systems Theory – EUROCAST 2017. LNCS, vol. 10672, pp. 20–28 (2018)
36. Zeifman, A.I., Korolev, V.Y., Satin, Y.A., Kiseleva, K.M.: Lower bounds for the rate of convergence for continuous-time inhomogeneous Markov chains with a finite state space. Stat. Probab. Lett. **137**, 84–90 (2018)
37. Zeifman, A., Satin, Y., Kiseleva, K., Korolev, V., Panfilova, T.: On limiting characteristics for a non-stationary two-processor heterogeneous system. Appl. Math. Comput. **351**, 48–65 (2019)

Convergence Rate Estimates for Some Models of Queuing Theory, and Their Applications

Alexander Zeifman, Yacov Satin, Anastasia Kryukova, Galina Shilova, and Ksenia Kiseleva

Abstract The forward Kolmogorov system for a general nonstationary Markovian queueing model with possible batch arrivals, possible catastrophes and state-dependent control at idle time is considered. We obtain upper bounds on the rate of convergence for corresponding models (nonstationary $M^X/M_n/1$ queue without catastrophes with the special resurrection intensities and general nonstationary $M^X/M_n/1$ queue with mass arrivals and catastrophes) and apply these estimates for some specific situations. Examples with given parameters are considered and corresponding plots are constructed.

1 Introduction

We consider forward Kolmogorov system for general nonstationary Markovian queueing model with possible batch arrivals, possible catastrophes and state-dependent control at idle time. The previous investigations in this area deal with different particular classes of this general model, see, for instance, [1–4, 6, 10]. Detailed discussion and references one can find in [4]. A general description of the model and basic results are given in [8]. Here we obtain upper bounds on the rate of convergence and apply them for some specific situations.

Let $X(t)$ be the queue-length process for this model. Denote by $\mathbf{p}(t)$ the column vector of state probabilities, $\mathbf{p}(t) = (p_0(t), p_1(t), \dots)^T$.

Then the probabilistic dynamics of the process $\{X(t),\ t \geq 0\}$ is given by the forward Kolmogorov system

A. Zeifman (✉)
Institute of Informatics Problems FRC CSC RAS, Vologda Research Center RAS,
Vologda State University, Lenina, 15, Vologda, Russia
e-mail: a_zeifman@mail.ru

Y. Satin · A. Kryukova · G. Shilova · K. Kiseleva
Vologda State University, Lenina, 15, Vologda, Russia

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_4

$$\frac{d\mathbf{p}(t)}{dt} = A(t)\mathbf{p}(t), \tag{1}$$

where

$$A(t) = \begin{pmatrix} q_{00}(t) & \beta_1(t)+\mu_1(t) & \beta_2(t) & \dots & \beta_j(t) & \dots & \\ h_1(t) & q_{11}(t) & \mu_2(t) & 0 & \dots & \dots & \dots \\ h_2(t) & b_1(t) & q_{22}(t) & \mu_3(t) & 0 & \dots & \dots \\ \vdots & \dots & \dots & \dots & \dots & \dots & \dots \\ \vdots & \dots & \dots & b_1(t) & q_{jj}(t) & \mu_{j+1}(t) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the transposed intensity matrix, with the following non-zero entries of $A(t)$:

$b_k(t)$ are the intensity of arrival of group of k customers to the non-empty queue, which does not depend on the current size of the length of queue;
$\mu_k(t)$ are the intensity of service of a customer in the queue, if the current size of the length of queue equals k;
$\beta_k(t)$ are the disaster (catastrophe) intensity, if the current size of the length of queue equals k;
$h_k(t)$ are the intensity of transition from zero to k (resurrection in terms of [4], or mass arrivals in terms of [1]);
$q_{kk}(t)$ are such that the corresponding column sums of $A(t)$ are zero for any $t \geq 0$.

Note that all "intensity" functions $b_k(t)$, $\mu_k(t)$, $\beta_k(t)$ and $h_k(t)$ are nonnegative for any $t \geq 0$, locally integrable on $[0, \infty)$, and bounded on this interval, namely, that $|q_{kk}(t)| \leq L < \infty$ for almost all $t \geq 0$.

Then, applying the modified combined approach of [5] and [7] we can obtain bounds on the rate of convergence of the queue-length process to its limiting characteristics and compute them. We separately consider the important special cases, see description in [9] and general results in [8].

Throughout the paper by $\|\cdot\|$ we denote the l_1-norm, i. e. $\|\mathbf{p}(t)\| = \sum_k |p_k(t)|$, and $\|A(t)\| = \sup_j \sum_i |a_{ij}|$. Let Ω be a set all stochastic vectors, i.e. l_1 vectors with non-negative coordinates and unit norm. Hence the operator function $A(t)$ from l_1 into itself is bounded for almost all $t \geq 0$ and locally integrable on $[0; \infty)$, moreover $\|A(t)\| = 2\sup_k |q_{kk}(t)| \leq 2L$ for almost all $t \geq 0$. Therefore we can consider (1) as a differential equation in the space l_1 with bounded operator, hence the Cauchy problem for differential Eq. (1) has a unique solutions for an arbitrary initial condition, and $\mathbf{p}(s) \in \Omega$ implies $\mathbf{p}(t) \in \Omega$ for $t \geq s \geq 0$.

Denote by $E(t, k) = E(X(t)|X(0) = k)$ the conditional expected number of customers in the system at instant t, provided that initially (at instant $t = 0$) k customers were present in the system.

Recall that a Markov chain $\{X(t),\ t \geq 0\}$ is called *weakly ergodic*, if $\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \to 0$ as $t \to \infty$ for any initial conditions $\mathbf{p}^*(0)$ and $\mathbf{p}^{**}(0)$, where $\mathbf{p}^*(t)$ and $\mathbf{p}^{**}(t)$ are the corresponding solutions of (1). A Markov chain $\{X(t),\ t \geq 0\}$ has the limiting mean $\varphi(t)$, if $\lim_{t\to\infty} (\varphi(t) - E(t,k)) = 0$ for any k.

2 Nonstationary $M^X/M_n/1$ Queue Without Catastrophes with the Special Resurrection Intensities

In this section we study as in [4] the queueing model without catastrophes (i.e. all $\beta_j(t) = 0$) with the special resurrection rates $h_j(t) = b_j(t)$, for any j, t. In addition, we suppose in this section that $b_{k+1}(t) \leq b_k(t)$ for all k.

In accordance with these assumptions, we arrive at the model described in [7] as queue with state-independent batch arrivals and state-dependent service intensities.

Let $\{d_i\}$ be a sequence of positive numbers, and D be an upper triangular matrix,

$$D = \begin{pmatrix} d_1 & d_1 & d_1 & \cdots \\ 0 & d_2 & d_2 & \cdots \\ 0 & 0 & d_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Denote by $\mathbf{y}(t) = D\mathbf{z}(t)$, where $\mathbf{z}(t) = \left(p_1^*(t) - p_1^{**}(t), p_2^*(t) - p_2^{**}(t), \ldots\right)^T$ is the difference of two solutions of the forward Kolmogorov system (1) with the corresponding initial conditions $\mathbf{p}^*(0)$ and $\mathbf{p}^{**}(0)$, in which all coordinates except p_0 are taken.

Put

$$\alpha_j(t) = \mu_j(t) - \frac{d_{j-1}}{d_j}\mu_{j-1}(t) + \sum_{i=1}^{\infty}\left(1 - \frac{d_{i+j}}{d_j}\right) b_i(t), \tag{2}$$

and

$$\alpha(t) = \inf \alpha_j(t). \tag{3}$$

Putting $p_0 = 1 - \sum_{i\geq 1} p_i$ and applying the logarithmic norm of operator function, see Theorem 1 in [8] and comparison of norms in [9], we get the following statement.

Proposition 1. *Let there exist an increasing sequence $\{d_j,\ j \geq 1\}$ of positive numbers with $d_1 = 1$, such that*

$$\int_0^\infty \alpha(t)\, dt = +\infty. \tag{4}$$

Then the Markov chain $X(t)$ is weakly ergodic and the following bound holds:

$$\|\mathbf{y}(t)\| \le e^{-\int_0^t \alpha(u)du}\|\mathbf{y}(0)\|, \tag{5}$$

and

$$\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\| \le 4e^{-\int_0^t \alpha(u)du}\|\mathbf{y}(0)\|, \tag{6}$$

for any initial conditions $\mathbf{p}^*(0)$, $\mathbf{p}^{**}(0)$ and any $t \ge 0$.

Moreover, if $W = \inf_{i\ge1}\frac{d_i}{i} > 0$, then $X(t)$ has the limiting mean and

$$|\varphi(t) - E(t,k)| \le \frac{2}{W}e^{-\int_0^t \alpha(u)du}\|\mathbf{y}(0)\|, \tag{7}$$

for any $t \ge 0$, and any k.

Here we apply all the bounds for nonstationary $M^X/M/S$ queue with batch arrivals and S servers, which is described and firstly studied in [7]. In this model we have the following intensities: $b_k(t) = \frac{1}{k}\lambda(t)$ if $1 \le k \le S$, and $b_k(t) = 0$ if $k > S$ are rates of arrival of a group of k customers; and $\mu_k(t) = \min(k, S)\mu(t)$ is the corresponding service rate.

Put $d_1 = 1$ and $d_{k+1} = \delta d_k$. Then

$$\alpha_k(t) = k\mu(t) - \frac{k-1}{\delta}\mu(t) + \sum_{i=1}^{S}\left(1-\delta^i\right)\lambda(t), \tag{8}$$

if $k \le S$, and

$$\alpha_k(t) = S\mu(t)\left(1-\frac{1}{\delta}\right) + \sum_{i=1}^{S}\left(1-\delta^i\right)\lambda(t), \tag{9}$$

if $k > S$.

Denote $\Delta = \left(1 + (\delta+1)/2 + \cdots + \left(\delta^{S-1} + \cdots + \delta^2 + \delta + 1\right)/S\right)$, then we obtain

$$\alpha(t) = \min\left(\alpha_1(t), \alpha_{S+1}(t)\right) = \mu(t)\min\left(1, S - \frac{S}{\delta}\right) - \Delta(\delta-1)\lambda(t). \tag{10}$$

Let now $\delta \in \left(1, \frac{S}{S-1}\right)$. Then $1 - \frac{1}{\delta} < \frac{1}{S}$ and hence

$$\alpha(t) = \left(1-\delta^{-1}\right)\left(S\mu(t) - \Delta\delta\lambda(t)\right). \tag{11}$$

Then all assumptions of Proposition 1 for queue-length process of $M^X/M/S$ queue are fulfilled if

$$\int_0^{\infty} (S\mu(t) - \Delta\delta\lambda(t))\, dt = +\infty. \tag{12}$$

Let now arrival and service rates be 1-periodic in time.
Denote by $\lambda^* = \int_0^1 \lambda(t)\,dt$ and by $\mu^* = \int_0^1 \mu(t)\,dt$.
If $\delta = 1$ then $\Delta = S$ and $S\mu(t) - \Delta\delta\lambda(t) = S\,(\mu(t) - \lambda(t))$.
Therefore, if $\mu^* > \lambda^*$ then $S\mu^* - \Delta\delta\lambda^* > 0$, if $\delta - 1 > 0$ is small enough.
Finally, in 1-periodic situation the assumptions of Proposition 1 hold if $\mu^* > \lambda^*$.

Example 1. Consider the $M^X/M/S$ queue with $S = 10$, $\mu(t) = \mu = 3$, $\lambda(t) = 1 + M \sin 2\pi\omega t$ and different values of amplitude M, frequency ω.

One can put $\delta = 1.1$, then $e^{-\int_0^t \alpha(u)du} \le 2e^{-t}$ and $W > 0.23$. Here all assumptions of Proposition 1 hold and one can obtain the corresponding bounds on the rate of convergence to the limiting characteristics. One of the most important of them is the mean number of customers in the queue (the mathematical expectation).

The limiting mathematical expectation of the process and its dependence on the amplitude and frequency of the intensity of the arrival of requirements is shown (Figs. 1, 2, 3, 4, 5 and 6).

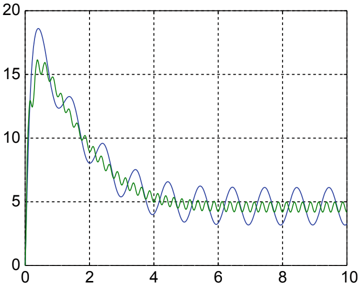


Fig. 1 Example 1. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 1, \omega = 1$ (blue) and $M = 1, \omega = 4$ (green)

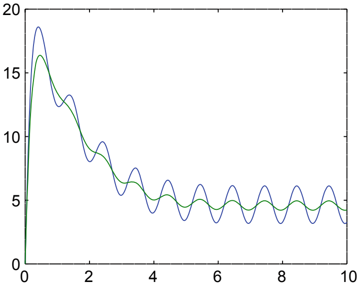


Fig. 2 Example 1. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 1, \omega = 1$ (blue) and $M = 0.25, \omega = 1$ (green)

3 General Nonstationary $M^X/M_n/1$ Queue with Mass Arrivals and Catastrophes

Consider here more general situation. Let resurrection intensities $h_j(t)$ be arbitrary locally integrable functions such that $h_0(t) = \sum_{i\geq 1} h_i(t) \leq L$ in accordance with our general assumptions.

Rewrite the forward Kolmogorov system (1) as

$$\frac{d\mathbf{p}}{dt} = A^*(t)\,\mathbf{p} + \mathbf{g}(t), \quad t \geq 0, \tag{13}$$

where $\mathbf{g}(t) = (\beta_*(t), 0, 0, \ldots)^T$, and $\beta_*(t) = \inf_i \beta_i(t)$. Then, applying the logarithmic norm of operator function, see Theorems 2 and 3 in [8], we get the following statements.

Proposition 2. *Let catastrophe rates be essential, i.e.*

$$\int_0^\infty \beta_*(t)\, dt = +\infty. \tag{14}$$

Then the queue-length process $X(t)$ is weakly ergodic in the uniform operator topology and the following bound holds

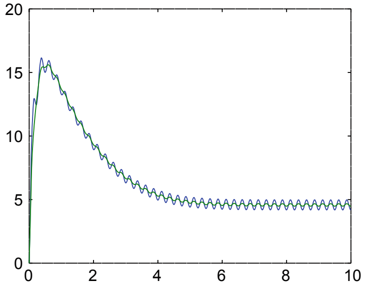


Fig. 3 Example 1. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 1$, $\omega = 4$ (blue) and $M = 0.25$, $\omega = 4$ (green)

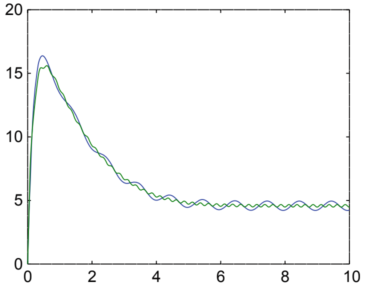


Fig. 4 Example 1. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 0.25$, $\omega = 1$ (blue) and $M = 0.25$, $\omega = 4$ (green)

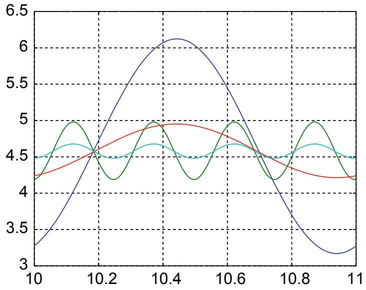


Fig. 5 Example 1. The mean $E(t, 0)$ for $t \in [10, 11]$ for all four cases

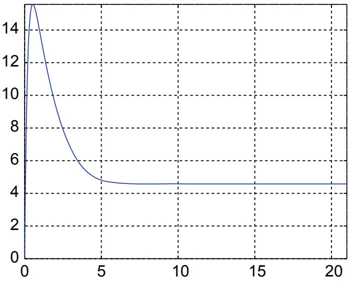


Fig. 6 Example 1. For comparison, the behaviour of the mean $E(t, 0)$ for the process with constant service rate ($M = 0$) is shown here

$$\left\|\mathbf{p}^*(t) - \mathbf{p}^{**}(t)\right\| \le e^{-\int\limits_0^t \beta_*(\tau)\,d\tau} \left\|\mathbf{p}^*(0) - \mathbf{p}^{**}(0)\right\| \le 2e^{-\int\limits_0^t \beta_*(\tau)\,d\tau}, \tag{15}$$

for any initial conditions $\mathbf{p}^*(0)$, $\mathbf{p}^{**}(0)$ *and any* $t \ge 0$.

Proposition 3. *Let* $\{d_i\}$, $1 = d_0 \le d_1 \le \ldots$ *be a non-decreasing sequence such that* $W = \inf_{i \ge 1} \frac{d_i}{i} > 0$, *and*

$$\int_0^\infty \beta_{**}(t)\,dt = +\infty, \tag{16}$$

where

$$\beta_{**}(t) = \inf_i \left(|a_{ii}^*(t)| - \sum_{j \ne i} \frac{d_j}{d_i} a_{ji}^*(t) \right), \tag{17}$$

and

$$a_{ij}^*(t) = \begin{cases} a_{0j}(t) - \beta_*(t), & \text{if } i = 0, \\ a_{ij}(t), & \text{otherwise}. \end{cases} \tag{18}$$

Then $X(t)$ has the limiting mean, say $\phi(t) = E(t, 0)$, and the following bound holds:

$$|E(t, j) - E(t, 0)| \leq \frac{1 + d_j}{W} e^{-\int\limits_0^t \beta_{**}(\tau)\, d\tau}, \tag{19}$$

for any j and any $t \geq 0$.

Now we apply this approach for nonstationary $M^X/M/S$ queue with batch arrivals, S servers, possible resurrections and catastrophes. The corresponding results for these models for some situations were firstly obtained in [5, 6].

Consider the model with the following intensities: $b_k(t) = \frac{1}{k}\lambda(t)$ if $1 \leq k \leq S$, $b_k(t) = 0$ if $k > S$ are rates of arrival of a group of k customers; $\mu_k(t) = \min(k, S)\mu(t)$ is the corresponding service rate. In addition, we consider only general restrictions on the intensity of resurrection and catastrophe, namely we suppose that resurrection rates are decreasing exponentially: $h_k(t) \leq cr^{-k}$ for some $r > 1$, $\beta_*(t) = \inf_i \beta_i(t)$.

Then the assumption of Proposition 2 is fulfilled if (14) hold.

Consider now the assumptions of Proposition 3. Put $d_0 = 1$ and $d_k = \delta^k$, where $\delta \in \left(1, \frac{S}{S-1}\right)$. Then we have $|a^*_{ii}(t)| - \sum_{j \neq i} \frac{d_j}{d_i} a^*_{ji}(t) \geq \beta_*(t) + \alpha(t)$, for $i \geq 1$, as in the previous Section. Let now $i = 0$. Then

$$|a^*_{00}(t)| - \sum_{j \neq 0} d_j a^*_{j0}(t) \geq \beta_*(t) - \sum_{k \geq 1} h_k(t) \left(\delta^k - 1\right) \geq \tag{20}$$

$$\beta_*(t) - c \sum_{k \geq 1} r^{-k} \left(\delta^k - 1\right) = \beta_*(t) - \frac{cr(\delta - 1)}{(r - \delta)(r - 1)}, \tag{21}$$

hence

$$\beta_{**}(t) \geq \beta_*(t) - \frac{cr(\delta - 1)}{(r - \delta)(r - 1)}, \tag{22}$$

and (16) implies the validity of all the assumptions of Proposition 3.

Let now intensities be 1-periodic in time. Denote $\lambda^* = \int_0^1 \lambda(t)\, dt$, $\mu^* = \int_0^1 \mu(t)\, dt$, $\beta^*_* = \int_0^1 \beta_*(t)\, dt$.

In this situation assumption of Proposition 2 hold if $\beta^*_* > 0$, and if, in addition, $\mu^* > \lambda^*$ then Proposition 3 is also true.

Example 2. Consider the $M^X/M/S$ queue with batch arrivals, S servers, resurrections and catastrophes with the following parameters: $S = 10$, $\mu(t) = \mu = 2$, $\lambda(t) = 1 + M \sin 2\pi\omega t$, $\beta_k(t) = \frac{1}{2} + \frac{1}{k+1}(1 + \sin 2\pi t)$; $h_k(t) = 2^{1-k}(1 + \cos 2\pi t)$.

One can put here $\delta = 1.02$, then $\beta_*(t) \geq 0.5$,

$$\beta_{**}(t) \geq \beta_*(t) - \frac{cr(\delta - 1)}{(r - \delta)(r - 1)} \geq 0.5 - \frac{8 \cdot 1.02}{0.98} \geq 0.3,$$

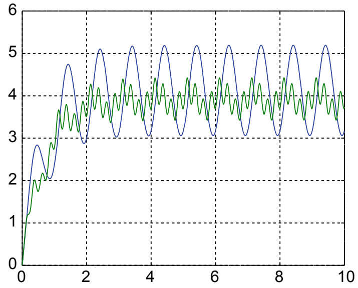


Fig. 7 Example 2. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 1, \omega = 1$ (blue) and $M = 1, \omega = 4$ (green)

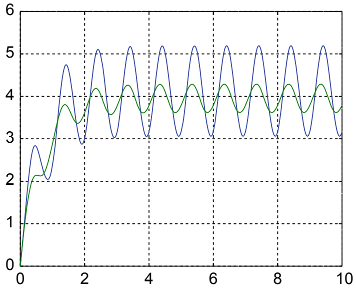


Fig. 8 Example 2. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 1, \omega = 1$ (blue) and $M = 0.25, \omega = 1$ (green).

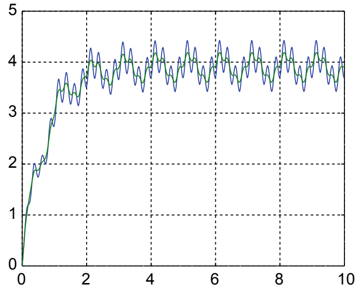


Fig. 9 Example 2. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 1, \omega = 4$ (blue) and $M = 0.25, \omega = 4$ (green)

and $W \geq 0.05$. Hence all assumptions of Propositions 2, 3 hold and one can obtain the corresponding bounds on the rate of convergence to the limiting characteristics. One of the most important of them is the mean number of customers in the queue (the mathematical expectation).

The limiting mathematical expectation of the process and its dependence on the amplitude and frequency of the intensity of the arrival of requirements is shown (Figs. 7, 8, 9, 10, 11 and 12).

Fig. 10 Example 2. The mean $E(t, 0)$ for $t \in [0, 10]$ with $M = 0.25$, $\omega = 1$ (blue) and $M = 0.25$, $\omega = 4$ (green)

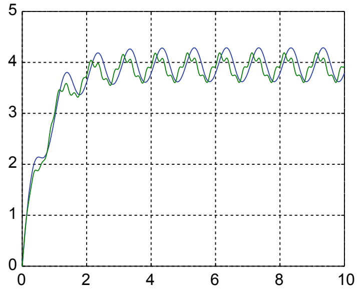


Fig. 11 Example 2. The mean $E(t, 0)$ for $t \in [10, 11]$ for all four cases

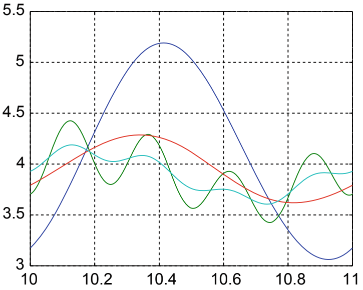
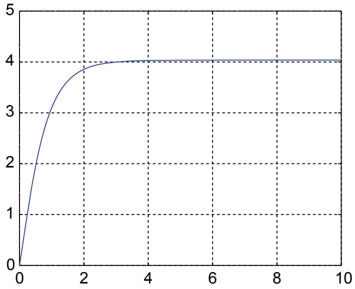


Fig. 12 Example 2. For comparison, the behaviour of the mean $E(t, 0)$ for the process with constant service rate ($M = 0$) is shown here



Acknowledgements The research has been supported by the Russian Science Foundation under grant 19-11-00020.

References

1. Chen, A., Renshaw, E.: The $M/M/1$ queue with mass exodus and mass arrivals when empty. J. Appl. Probab. **34**(1), 192–207 (1997)
2. Chen, A., Renshaw, E.: Markovian bulk-arriving queues with state-dependent control at idle time. Adv. Appl. Probab. **36**(2), 499–524 (2004)
3. Di Crescenzo, A., Giorno, V., Nobile, A.G., Ricciardi, L.M.: A note on birth-death processes with catastrophes. Stat. Probab. Lett. **78**(14), 2248–2257 (2008)

4. Li, J., Zhang, L.: $M^X/M/c$ Queue with catastrophes and state-dependent control at idle time. Front. Math. China. **12**(6), 1427–1439 (2017)
5. Zeifman, A., Korotysheva, A., Satin, Y., Kiseleva, K., Korolev, V., Shorgin, S.: Bounds for Markovian queues with possible catastrophes. In: Proceedings of 31st European Conference on Modelling and Simulation ECMS 2017, Digitaldruck Pirrot GmbHP Dudweiler, Germany, pp. 628–634 (2017)
6. Zeifman, A., Korotysheva, A., Satin, Y., Razumchik, R., Korolev, V., Shorgin, S.: Ergodicity and truncation bounds for inhomogeneous birth and death processes with additional transitions from and to origin. Stochast. Models **33**, 598–616 (2017)
7. Zeifman, A., Razumchik, R., Satin, Y., Kiseleva, K., Korotysheva, A., Korolev, V.: Bounds on the rate of convergence for one class of inhomogeneous Markovian queueing models with possible batch arrivals and services. Int. J. Appl. Math. Comput. Sci. **28**(1), 141–154 (2018)
8. Zeifman, A., Satin, Y., Kiseleva, K., Panfilova, T., Kryukova, A., Shilova, G., Sipin, A., Fokicheva, E.: Bounds on the rate of convergence for nonstationary $M^X/M_n/1$ queue with catastrophes and state-dependent control at idle time. In: International Conference on Computer Aided Systems Theory. LNCS, vol. 12013, pp. 143–149 (2020)
9. Zeifman, A.: On the study of forward Kolmogorov system and the corresponding problems for inhomogeneous continuous-time Markov chains (2020, this issue)
10. Zhang, L., Li, J.: The M/M/c queue with mass exodus and mass arrivals when empty. J. Appl. Probab. **52**, 990–1002 (2015)

On Enlarged Sufficient Conditions for L^2-Dissipativity of Linearized Explicit Schemes with Regularization for 1D Gas Dynamics Systems of Equations

Alexander Zlotnik

Abstract We study an explicit two-level in time and three-point symmetric in space finite-difference scheme for 1D barotropic and full gas dynamics systems of equations. The scheme is a linearization at a constant background solution (with an arbitrary velocity) of finite-difference schemes with general viscous regularization. We enlarge recently proved sufficient conditions (on the Courant-like number) for L^2-dissipativity in the Cauchy problem for the schemes by deriving new bounds for the commutator of matrices of viscous and convective terms. We deal with the case of a kinetic regularization in more detail and specify sufficient conditions in this case where the mentioned matrices are closely connected. Importantly, these new sufficient conditions rapidly tend to the known necessary ones as the Mach number grows. Also several forms of setting a regularization parameter are considered.

Keywords 1D gas dynamics equations · Viscous regularization · Explicit finite-difference scheme · L^2-dissipativity · Commutator of matrices of viscous and convective terms · Mach number

1 Introduction

Vast literature is devoted to numerical methods for solving gas dynamics systems of equations, e.g., see [1, 6, 7, 12]. Among them, a class of finite-difference and finite volume methods based on regularizations of such systems exist. For high-performance computing, explicit in time methods are especially convenient. An important practical question concerns stability conditions for explicit methods.

In the linearized at constant solution statement, explicit finite-difference methods based on regularizations contain matrices A of viscous terms and B of convective ones. A criterion as well as necessary and sufficient conditions for L^2-dissipativity in the Cauchy problem for explicit two-level in time and three-point symmetric in

A. Zlotnik (✉)
Department of Mathematics at Faculty of Economic Sciences, National Research University Higher School of Economics, Pokrovskii bd. 11, 109028 Moscow, Russia
e-mail: azlotnik@hse.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_5

space finite-difference schemes have recently been given for 1D gas dynamics systems of equations in [16, 17] by the spectral method [3, 8]. The criterion contains the commutator of A and B, and its treating leads to the necessary and sufficient conditions. Both full and simpler barotropic gas dynamics systems have been considered; recall that the barotropic system also finds various practical applications. The case of a kinetic, or quasi-gasdynamic (QGD), regularization [2, 5, 9] has been studied in more detail. The practical relevance of the found conditions in the nonlinear statement has been analyzed in [17, 18]. Some more rough sufficient conditions in particular cases were given previously in [9, 10]. Recall that the Petrovskii parabolicity of the QGD systems, stability of small perturbations and energy relations for them were studied in [13, 14] confirming the regularization properties of these systems. Note that other recent regularizations were suggested in [4, 11].

In this paper, enlarged sufficient Courant-type stability conditions are derived for the same linearized schemes once again. This is achieved by means of applying the new improved bound for the commutator of A and B in the L^2-dissipativity criterion. The case of the QGD-regularization is considered in more detail both for the barotropic and full 1D gas dynamics systems where the specific relations between A and B (in particular, the independence of their commutator from the Mach number) is essentially used. Importantly, the new sufficient stability conditions rapidly tend to the known sufficient conditions as the Mach number grows. This is essential in view of computing super- and hypersonic flows.

In addition, several forms of choosing a regularization parameter τ are considered including its dependence on the spatial step and the sound velocity, without or with the dependence on the gas velocity (or the Mach number), and dependence on the time step only.

The paper is organized as follows. In Sect. 2, the criterion and necessary and sufficient conditions for L^2-dissipativity of an abstract (using general matrices A and B) explicit two-level in time and three-point symmetric in space 1D scheme with a regularization from [16, 17] are recalled, and a new enlarged sufficient condition together with bounds for some particular commutator-like matrices are proved. In Sects. 3 and 4, similar results are specified in the case of the QGD regularization for the 1D barotropic and full gas dynamics systems, respectively. Several forms of choosing the regularization parameter τ are covered as well.

2 Conditions for L^2-Dissipativity of an Abstract Explicit Scheme with a Regularization

The linearized at the constant solution 1D regularized gas dynamics systems of equations can be written in the vector form

$$\partial_t \mathbf{z} + B\partial_x \mathbf{z} - \tau_* c_*^2 A \partial_x^2 \mathbf{z} = 0, \tag{1}$$

where $\mathbf{z} = \mathbf{z}(x,t)$ is a $\mathbb{C}^n$-valued function defined for $x \in \mathbb{R}$ and $t \geqslant 0$, $A = A^*$ and $B = B^*$ are Hermitian matrices (of viscous and convective terms) of order n, $c_* > 0$ is a scaling factor (the characteristic velocity) and $\tau_* > 0$ is the regularization parameter. Hereafter ∂_t and ∂_x are the partial derivatives in time $t \geqslant 0$ and $x \in \mathbb{R}$. For the kinetic regularization, see more details in Sects. 3 and 4 below and [14].

It is not difficult to see that, in the case $A \geqslant 0$, i.e. $(A\xi, \xi)_{\mathbb{C}^n} \geqslant 0$ for any $\xi \in \mathbb{C}^n$, the solution to Eq. (1) supplemented with the initial condition $\mathbf{z}|_{t=0} = \mathbf{z}_0$ (i.e., of the Cauchy problem) satisfies the bounds

$$\sup_{t\geqslant 0} \||\mathbf{z}(\cdot,t)|\|_{L^2(\mathbb{R})} \leqslant \||\mathbf{z}_0|\|_{L^2(\mathbb{R})} \tag{2}$$

and $\sqrt{\tau_*}c_*\|(A\partial_x\mathbf{z}, \partial_x\mathbf{z})_{\mathbb{C}^n}\|_{L^2(\mathbb{R}\times(0,+\infty))} \leqslant \sqrt{2}\||\mathbf{z}_0|\|_{L^2(\mathbb{R})}$.

We define the uniform mesh ω_h on $\mathbb{R}$ with the nodes $x_k = kh$, $k \in \mathbb{Z}$, and the step $h > 0$. Let $\bar{\omega}^{\Delta t}$ be the uniform mesh in t with the nodes $t_m = m\Delta t$, $m \geqslant 0$, and the step $\Delta t > 0$. We set $v_k^m = v(x_k, t_m)$ and define the finite-difference mesh operators in x and t

$$\mathring{\delta} v_k = \frac{v_{k+1} - v_{k-1}}{2h}, \quad \delta^*\delta v_k = \frac{v_{k+1} - 2v_k + v_{k-1}}{h^2}, \quad \delta_t v^m = \frac{v^{m+1} - v^m}{\Delta t}, \quad v^{+,m} = v^{m+1}.$$

Let H be the Hilbert space of $\mathbb{C}^n$-valued vector functions defined and square summable on ω_h and endowed with the inner product $(\mathbf{v}, \mathbf{y})_H = h \sum_{k=-\infty}^{+\infty} (\mathbf{v}_k, \mathbf{y}_k)_{\mathbb{C}^n}$.

In this section, we consider an abstract explicit linear finite-difference scheme that is two-level in time and symmetric three-point in space for the linearized system of Eq. (1)

$$\delta_t \mathbf{y} + c_* B\mathring{\delta}\mathbf{y} - \tau_* c_*^2 A\delta^*\delta\mathbf{y} = 0 \ \text{ on } \ \omega_h \times \bar{\omega}^{\Delta t}, \tag{3}$$

where $\mathbf{y}^m \in H$ for $m \geqslant 0$. Notice that in the particular case $A = B^2$ and $\tau_* = \frac{\Delta t}{2}$ the stability analysis of the similar linearized Lax-Wendroff scheme is given in [8].

We are interested in the conditions for validity of the uniform in time bound

$$\sup_{m\geqslant 0} \|\mathbf{y}^m\|_H \leqslant \|\mathbf{y}^0\|_H \ \ \forall\, \mathbf{y}^0 \in H \tag{4}$$

which is the finite-difference counterpart of bound (2). Scheme (3) can be rewritten in the explicit form

$$\mathbf{y}^+ = \mathscr{A}\mathbf{y} \equiv \mathbf{y} - \Delta t c_0 B\mathring{\delta}\mathbf{y} + \Delta t\tau_* c_*^2 A\delta^*\delta\mathbf{y},$$

where $\mathscr{A}\colon H \to H$. Recall that bound (4) is equivalent to bound $\|\mathscr{A}\| \leqslant 1$ or the H-dissipativity property

$$\|\mathbf{y}^m\|_H \leqslant \|\mathbf{y}^{m-1}\|_H \leqslant \ldots \leqslant \|\mathbf{y}^0\|_H \ \ \forall\, \mathbf{y}^0 \in H, \ \ \forall\, m \geqslant 1.$$

Let first $\varDelta t$ and τ_* be given by the formulas [2, 5]

$$\varDelta t = \tilde{\beta}\frac{h}{c_*},\quad \tau_* = \alpha\frac{h}{c_*} \tag{5}$$

with the parameters $\tilde{\beta} > 0$ (the Courant-like number) and $\alpha > 0$. We are interested in conditions on $\tilde{\beta}$ depending on α such that bound (4) holds.

We first recall the matrix criterion together with necessary and sufficient conditions for that proved by the spectral method [16, 17]. Let $[A, B] = AB - BA$ be the commutator of the matrices A and B. Recall that, for Hermitian matrices A and B, $\mathbf{i}[A, B]$ is a Hermitian matrix too, where $\mathbf{i}$ is the imaginary unit.

Theorem 1. *1. Bound* (4) *is valid if and only if the matrix inequality holds*

$$\tilde{\beta}\Big(2\sigma\alpha A^2 + \frac{1-\sigma}{2\alpha}B^2 \pm \sqrt{\sigma(1-\sigma)}\mathbf{i}[A, B]\Big) \leqslant A \ \ \forall 0 < \sigma \leqslant 1. \tag{6}$$

2. The matrix inequalities

$$2\alpha\tilde{\beta}A \leqslant I,\quad \frac{\tilde{\beta}}{2\alpha}B^2 \leqslant A \tag{7}$$

are necessary and the matrix inequality

$$\tilde{\beta}\Big(2\alpha A^2 + \frac{1}{2\alpha}B^2\Big) \leqslant A \tag{8}$$

is sufficient for the validity of bound (4).

The second inequality (7) implies that $A \geqslant 0$. Recall that the derivation of the sufficient condition (8) is based on the rather rough bound

$$\pm\mathbf{i}[A, B] \leqslant \varepsilon A^2 + \varepsilon^{-1}B^2 \quad \forall\varepsilon > 0.$$

Now we present a new general sufficient condition for the validity of bound (4) provided that another bound for $\pm\mathbf{i}[A, B]$ is available. Denote by $\lambda_{\max}(A)$ the maximal eigenvalue of a Hermitian matrix A.

Theorem 2. *Let* $A \geqslant 0$ *and the matrix inequalities* $B^2 \leqslant c_B A$ *and* $\pm\mathbf{i}[A, B] \leqslant c_A A$ *be valid with some* $c_B \geqslant 0$ *and* $c_A \geqslant 0$. *Then bound* (4) *holds under the condition*

$$\tilde{\beta}\Big\{\alpha\lambda_{\max}(A) + \frac{c_B}{4\alpha} + \Big[\Big(\alpha\lambda_{\max}(A) - \frac{c_B}{4\alpha}\Big)^2 + \frac{c_A^2}{4}\Big]^{1/2}\Big\} \leqslant 1. \tag{9}$$

Proof. The inequality $A^2 \leqslant \lambda_{\max}(A)A$ and those from the hypotheses imply that criterion (6) holds under the condition

$$\tilde{\beta}\varphi(\sigma) \leqslant 1 \ \ \forall 0 < \sigma \leqslant 1,\ \text{ with } \varphi(\sigma) := a\sigma + b(1-\sigma) + c_A\sqrt{\sigma(1-\sigma)} \text{ on } [0, 1],$$

where $a := 2\alpha\lambda_{\max}(A) \geqslant 0$ and $b := \frac{c_B}{2\alpha} \geqslant 0$. Obviously we have

$$\varphi'(\sigma) = a - b + c_A \frac{1-2\sigma}{2\sqrt{\sigma(1-\sigma)}} \ \text{ on } \ (0,1),$$

and then

$$\varphi'(\sigma_0) = 0 \ \text{ for } \ \sigma_0 = \frac{1}{2} + \frac{a-b}{2\sqrt{(a-b)^2 + c_A^2}} \in (0,1) \ \text{ for } \ c_A > 0.$$

Then it is straightforward to check that

$$\max_{0 \leqslant \sigma \leqslant 1} \varphi(\sigma) = \varphi(\sigma_0) = \frac{1}{2}(a+b) + \frac{1}{2}\sqrt{(a-b)^2 + c_A^2}$$

for $c_A \geqslant 0$ and

$$\max\{a, b\} \leqslant \varphi(\sigma_0) \leqslant \max\{a, b\} + \frac{c_A}{2}.$$

Also $\varphi(\sigma_0) < a + b$ (or $\varphi(\sigma_0) = a + b$) is equivalent to $c_A^2 < 4ab$ (or $c_A^2 = 4ab$). These relations lead to (9) and Remark 1.

Remark 1. The following two-sided bound for the term in the curly brackets in (9) holds

$$\max\left\{2\alpha\lambda_{\max}(A), \frac{c_B}{2\alpha}\right\} \leqslant \alpha\lambda_{\max}(A) + \frac{c_B}{4\alpha} + \left[\left(\alpha\lambda_{\max}(A) - \frac{c_B}{4\alpha}\right)^2 + \frac{c_A^2}{4}\right]^{1/2}$$
$$\leqslant \max\left\{2\alpha\lambda_{\max}(A), \frac{c_B}{2\alpha}\right\} + \frac{c_A}{2}. \tag{10}$$

The relation

$$\alpha\lambda_{\max}(A) + \frac{c_B}{4\alpha} + \left[\left(\alpha\lambda_{\max}(A) - \frac{c_B}{4\alpha}\right)^2 + \frac{c_A^2}{4}\right]^{1/2}$$
$$< 2\lambda_{\max}(A) + \frac{c_B}{2\alpha} \left(\text{or } = 2\lambda_{\max}(A) + \frac{c_B}{2\alpha}\right)$$

is equivalent to $c_A < 4c_B\lambda_{\max}(A)$ (or $c_A = 4c_B\lambda_{\max}(A)$) for any $\alpha > 0$.

Remark 2. The optimal value $\alpha = \alpha_{opt}$ in (9) is such that the expression in the curly brackets is minimal. Calculating the derivative of the expression, one can see that it is zero only for $\alpha = \alpha_{opt} = \frac{1}{2}\sqrt{\frac{c_B}{\lambda_{\max}(A)}}$, and the mentioned minimal value equals $\sqrt{c_B\lambda_{\max}(A)} + \frac{1}{2}c_A$.

On the other hand, in the nonlinear statement, this value is not always the best one so that an analysis involving other values of α is also of interest.

The next algebraic lemma is crucial below when applying Theorem 2.

Lemma 1. *Let $A = A^T > 0$ be a real matrix of order $n = 2, 3$ and $a, b \in \mathbb{R}$. The following matrix inequalities hold:*

$$\pm\mathbf{i}\begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} \leqslant c_A A \ \ with\ \ c_A := \frac{|a|}{\sqrt{|A|}} \tag{11}$$

for $n = 2$ and

$$\pm\mathbf{i}\begin{pmatrix} 0 & a & 0 \\ -a & 0 & b \\ 0 & -b & 0 \end{pmatrix} \leqslant c_A A \ \ with\ \ c_A := \max\left\{\frac{|a|}{\sqrt{|A_{12}|}}, \frac{|b|}{\sqrt{|A_{23}|}}, \sqrt{\frac{\mathbf{b}^T A_{13}\mathbf{b}}{|A|}}\right\}, \tag{12}$$

where

$$A_{kl} = \begin{pmatrix} a_{kk} & a_{kl} \\ a_{kl} & a_{ll} \end{pmatrix} \ \ for\ \ k < l,\ \ \mathbf{b} = \begin{pmatrix} b \\ a \end{pmatrix}, \tag{13}$$

for n = 3. Here, for example, $|A|$ is the determinant of A.

Proof. Denote by C the matrices on the left in inequalities (11) and (12) (without $\pm$). Then

$$tC + A = \begin{pmatrix} a_{11} & a_{12} + \mathbf{i}ta \\ a_{12} - \mathbf{i}ta & a_{22} \end{pmatrix} \text{ for } n = 2,$$

$$tC + A = \begin{pmatrix} a_{11} & a_{12} + \mathbf{i}ta & a_{13} \\ a_{12} - \mathbf{i}ta & a_{22} & a_{23} + \mathbf{i}tb \\ a_{13} & a_{23} - \mathbf{i}tb & a_{33} \end{pmatrix} \text{ for } n = 3,$$

where t is a real parameter, and these inequalities are equivalent to the property $tC + A \geqslant 0$ for $t = \pm\frac{1}{c_A}$ and $c_A > 0$ (for $c_A = 0$, the inequalities become trivial).

According to the Sylvester-type criterion, this property is valid if and only if all the principal minors of $tC + A$ are non-negative. Since $A > 0$, this means validity of a unique condition

$$|tC + A| = a_{11}a_{22} - (a_{12}^2 + t^2a^2) = |A| - t^2a^2 \geqslant 0 \ \text{ for } \ n = 2$$

or three conditions

$$|(tC + A)_{12}| = |A_{12}| - t^2a^2 \geqslant 0,\ \ |(tC + A)_{23}| = |A_{23}| - t^2b^2 \geqslant 0,$$
$$|(tC + A)| = a_{11}a_{22}a_{33} + 2a_{13}\,\mathrm{Re}\left[(a_{12} + \mathbf{i}ta)(a_{23} + \mathbf{i}tb)\right] - a_{13}^2a_{22}$$
$$-(a_{12}^2 + t^2a^2)a_{33} - a_{11}(a_{23}^2 + t^2b^2) = |A| - t^2\mathbf{b}^T A_{13}\mathbf{b} \geqslant 0 \ \text{ for } \ n = 3. \tag{14}$$

Clearly these inequalities lead to the result, and the indicated value of c_A is the minimal possible.

Remark 3. For $n = 3$, Lemma 1 remains valid in the case $A \geqslant 0$, $|A| = 0$ together with $|A_{12}| > 0$, $|A_{23}| > 0$, $\mathbf{b}^T A_{13}\mathbf{b} = 0$ and c_A with the omitted third term in (12). Here inequality (14) is valid automatically.

3 An Enlarged Sufficient Condition for L^2-Dissipativity in the 1D Barotropic Case

The barotropic quasi-gasdynamic (QGD) system of equations [13, 14] in the 1D case consists of the regularized mass and momentum balance equations

$$\partial_t\rho + \partial_x j = 0, \quad \partial_t(\rho u) + \partial_x\big(ju + p(\rho) - \Pi\big) = 0. \tag{15}$$

The sought functions $\rho > 0$ and u are the gas density and velocity, as well as $p = p(\rho)$ is the pressure with $p'(\rho) > 0$. Also j and Π are the regularized mass flux and viscous stress given by

$$j = \rho(u - w), \quad w = \frac{\tau}{\rho}u\partial_x(\rho u) + \widehat{w}, \quad \widehat{w} = \frac{\tau}{\rho}\big(\rho u\partial_x u + \partial_x p(\rho)\big), \tag{16}$$

$$\Pi = \Pi_{NS} + \rho u\widehat{w} + \tau p'(\rho)\partial_x(\rho u), \quad \Pi_{NS} = \mu\partial_x u, \tag{17}$$

where w and $\widehat{w}$ are the regularizing velocities, $\tau = \tau(\rho, u) > 0$ is a regularization parameter and Π_{NS} is the Navier-Stokes-type viscous stress with the viscosity coefficient $\mu \geqslant 0$. The artificial viscosity coefficient μ is given by the standard QGD-formula $\mu = \alpha_S\tau\rho p'(\rho)$ with the parameter $\alpha_S \geqslant 0$ (the Schmidt number).

The barotropic QGD system is simplified to the barotropic compressible Navier-Stokes system, or the barotropic Euler one, for respectively $\tau = 0$ and $\mu > 0$, or $\tau = \mu = 0$.

System (15)–(17) can be linearized [14] at a constant background solution $\rho_* > 0$ and u_* by setting $\rho = \rho_* + \rho_*\tilde{\rho}$ and $u = u_* + c_*\tilde{u}$, where $c_* = \sqrt{p'(\rho_*)}$ is the sound velocity, and taking $\tau_* = \tau(\rho_*, u_*)$. The linearized system for the scaled small perturbations $\mathbf{z} = (\tilde{\rho}, \tilde{u})^T$ can be written in the vector form (1) with the matrices

$$B = \begin{pmatrix} M & 1 \\ 1 & M \end{pmatrix}, \quad A = \begin{pmatrix} M^2 + 1 & 2M \\ 2M & \alpha_S + M^2 + 1 \end{pmatrix}. \tag{18}$$

Hereafter $M := \frac{u_*}{c_*}$, and $|M|$ is the Mach number. The result of the linearization of some schemes for the barotropic QGD system coincides with (3) with matrices (18), see [17].

Recall the result on the necessary and sufficient conditions for L^2-dissipativity from [17] (in the modified form).

Theorem 3. *For scheme* (3) *with matrices* (18) *($n=2$), the necessary condition* (7) *takes the form*

$$\tilde{\beta}\max\left\{2\alpha\lambda_{\max}(A),\,\frac{1}{2\alpha}\right\}\leqslant 1, \tag{19}$$

and the sufficient condition (8) *is valid for*

$$\tilde{\beta}\Big(2\alpha\lambda_{\max}(A)+\frac{1}{2\alpha}\Big)\leqslant 1. \tag{20}$$

Here $\lambda_{\max}(A)=\frac{\alpha_S}{2}+M^2+1+\sqrt{\left(\frac{\alpha_S}{2}\right)^2+4M^2}$.

Let us give a new enlarged sufficient condition for the validity of bound (4).

Theorem 4. *For scheme* (3) *with matrices* (18) *($n=2$), bound* (4) *is valid under the condition*

$$\tilde{\beta}\left\{\alpha\lambda_{\max}(A)+\frac{1}{4\alpha}+\left[\Big(\alpha\lambda_{\max}(A)-\frac{1}{4\alpha}\Big)^2+\frac{c_A^2}{4}\right]^{1/2}\right\}\leqslant 1 \tag{21}$$

where $c_A^2=\frac{\alpha_S^2}{|A|}$ *in the case* $\alpha_S>0$ *or* $|M|\neq 1$*; otherwise* $c_A=0$*. Here* $|A|=(M^2-1)^2+\alpha_S(M^2+1)$.

Proof. It is known [17] and easy to check that

$$A=B^2+D,\quad D=\operatorname{diag}\{0,\alpha_S\}\geqslant 0;$$

hereafter diag{...} is a diagonal matrix with the listed diagonal elements. Thus $A\geqslant 0$ and $B^2\leqslant A$ (i.e., $c_B=1$). Note that $|A|=0$ if and only if $\alpha_S=0$ and $|M|=1$; otherwise $A>0$. Therefore it is straightforward to check that

$$[A,B]=[D,B]=\begin{pmatrix}0 & -\alpha_S\\ \alpha_S & 0\end{pmatrix},$$

and owing to inequality (11) we have $\pm\mathbf{i}[A,B]\leqslant c_A A$ with $c_A=\frac{\alpha_S}{|A|^{1/2}}$ for $\alpha_S>0$ or $|M|\neq 1$. For $\alpha_S=0$, simply $[A,B]=0$ and thus $c_A=0$. Now Theorem 2 implies the result.

We have $\lambda_{\max}(A)\geqslant\alpha_S+1$ and $|A|\geqslant\alpha_S$, therefore $c_A<\lambda_{\max}(A)$ and according to Remark 1, the sufficient condition (21) is broader than (20) for any $\alpha_S\geqslant 0$ (for $\alpha_S=0$ this is obvious). For $c_A=0$, the sufficient condition (21) coincides with the necessary one (19) and becomes a criterion.

In addition, $c_A=O\big(\frac{1}{M^2}\big)$ as $M\to\infty$ and, moreover, the sufficient condition (21) rapidly tends to the necessary one (19) as $|M|$ grows.

For significant Mach numbers, another form of formulas (5) is preferable

$$\Delta t = \beta \frac{h}{|u_*| + c_*}, \quad \tau_* = \widehat{\alpha} \frac{h}{|u_*| + c_*} \tag{22}$$

with $\beta = \tilde{\beta}(|M| + 1)$ and $\widehat{\alpha} = \alpha(|M| + 1)$, see [16, 17]. In terms of β and $\widehat{\alpha}$, condition (21) can be rewritten in the form

$$\beta\left\{\widehat{\alpha}\frac{\lambda_{\max}(A)}{(|M|+1)^2} + \frac{1}{4\widehat{\alpha}} + \left[\left(\widehat{\alpha}\frac{\lambda_{\max}(A)}{(|M|+1)^2} - \frac{1}{4\widehat{\alpha}}\right)^2 + \frac{c_A^2}{4(|M|+1)^2}\right]^{1/2}\right\} \leqslant 1. \tag{23}$$

Its advantage is that the optimal value $\widehat{\alpha} = \widehat{\alpha}_{opt} \equiv \frac{|M|+1}{2\sqrt{\lambda_{\max}(A)}}$ (as well as the corresponding minimal value $\frac{\sqrt{\lambda_{\max}(A)}+\frac{1}{2}c_A}{|M|+1}$ of the expression in the curly brackets, see Remark 2) varies weakly with respect to the Mach number. And this result is an advantage of the kinetic regularization itself.

One can rewrite the new enlarged sufficient condition (23) in the form $\beta \leqslant \beta_{\text{suf}}(\widehat{\alpha}, M)$ where $\beta_{\text{suf}}(\widehat{\alpha}, M)$ is the inverse of the expression in the curly brackets. Let us also rewrite the necessary condition (19) and the previous sufficient condition (20) in the similar forms $\beta \leqslant \beta_{\text{nec}}(\widehat{\alpha}, M)$ and $\beta \leqslant \beta_{\text{suf}}^{(0)}(\widehat{\alpha}, M)$, respectively. In Fig. 1, we present graphs of functions $\beta_{\text{nec}}(\widehat{\alpha}, M)$ (solid line), $\beta_{\text{suf}}(\widehat{\alpha}, M)$ (dot line) and $\beta_{\text{suf}}^{(0)}(\widehat{\alpha}, M)$ (dotted line) for $M = 0, 1, 1.5$ and 2 in the typical case $\alpha_S = 1$. Clearly $\beta_{\text{suf}}(\widehat{\alpha}, M)$ is much closer to $\beta_{\text{nec}}(\widehat{\alpha}, M)$ than $\beta_{\text{suf}}^{(0)}(\widehat{\alpha}, M)$ for all these M and, moreover, $\beta_{\text{suf}}(\widehat{\alpha}, M)$ rapidly tends to $\beta_{\text{nec}}(\widehat{\alpha}, M)$ as the Mach number M grows; in particular, for $M = 2$, they are already practically identical (though actually they very slightly differ near their maximums).

Let us also discuss an alternative choice $\tau_* = \frac{a_0}{2}\Delta t$ *independent of* h. In this case, the above formulas for Δt are not required any more. We can cover this case simply by the formal change $\alpha = \frac{a_0 c_* \Delta t}{2h}$ (that is possible since above $\alpha > 0$ was arbitrary) and setting $\tilde{\beta} = \frac{c_* \Delta t}{h}$. Then the necessary condition (19) takes the form

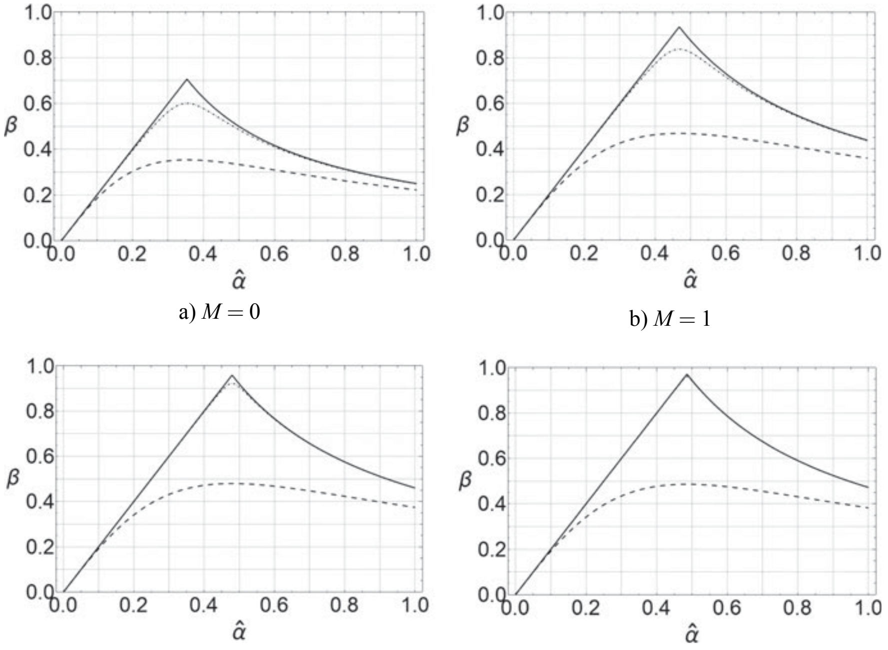
$$\frac{c_* \Delta t}{h} \max\left\{\frac{a_0 c_* \Delta t}{2h}\lambda_{\max}(A), \frac{h}{a_0 c_* \Delta t}\right\} \leqslant 1,$$

or, equivalently, the Courant-like form

$$\sqrt{a_0 \lambda_{\max}(A)}\,\frac{c_* \Delta t}{h} \leqslant 1, \quad a_0 \geqslant 1. \tag{24}$$

The sufficient condition (20) is converted into the conditions

$$\frac{a_0}{\sqrt{a_0 - 1}}\sqrt{\lambda_{\max}(A)}\,\frac{c_* \Delta t}{h} \leqslant 1, \quad a_0 > 1. \tag{25}$$



a) $M = 0$

b) $M = 1$

c) $M = 1.5$

d) $M = 2$

Fig. 1 Graphs of functions $\beta_{\mathrm{nec}}(\widehat{\alpha}, M)$ (solid line), $\beta_{\mathrm{suf}}(\widehat{\alpha}, M)$ (dot line) and $\beta \leqslant \beta_{\mathrm{suf}}^{(0)}(\widehat{\alpha}, M)$ (dotted line) for $M = 0, 1, 1.5$ and 2 in the case $\alpha_S = 1$

Concerning the new sufficient condition (21), it first can be rewritten as

$$\frac{1}{2}a_0\lambda_{\max}(A)\tilde{\beta}^2 + \frac{1}{2a_0} + \left[\left(\frac{1}{2}a_0\lambda_{\max}(A)\tilde{\beta}^2 - \frac{1}{2a_0}\right)^2 + \frac{c_A^2\tilde{\beta}^2}{4}\right]^{1/2} \leqslant 1$$

since $\alpha = \frac{1}{2}a_0\tilde{\beta}$. Similarly to (10), we get $\frac{1}{a_0} \leqslant 1$. If the condition $\frac{1}{2}a_0\lambda_{\max}(A)\tilde{\beta}^2 + \frac{1}{2a_0} \leqslant 1$ is valid, we can transfer the first and second terms on the left to the right, square the both sides and, after reducing similar terms, derive

$$\left((a_0 - 1)\lambda_{\max}(A) + \frac{c_A^2}{4}\right)\tilde{\beta}^2 \leqslant 1 - \frac{1}{a_0}.$$

Note that this inequality implies the last mentioned condition. Therefore finally we transform (21) into the form

$$\sqrt{a_0\lambda_{\max}(A) + \frac{a_0}{a_0 - 1}\frac{c_A^2}{4}}\,\frac{c_*\Delta t}{h} \leqslant 1, \quad a_0 > 1. \tag{26}$$

4 An Enlarged Sufficient Condition for L^2-Dissipativity in the Case of the Full System

The full QGD system of equations [2, 5] in the 1D case consists of the following mass, momentum and total energy balance equations

$$\partial_t \rho + \partial_x j = 0, \quad \partial_t(\rho u) + \partial_x(ju + p - \Pi) = 0, \tag{27}$$

$$\partial_t E + \partial_x\left[(E + p)(u - w)\right] = \partial_x(-q + \Pi u). \tag{28}$$

The function $E = 0.5\rho u^2 + \rho\varepsilon$ is the total energy, and $\varepsilon > 0$ is the specific internal energy. We consider the perfect polytropic gas state equation $p = (\gamma - 1)\rho\varepsilon$ with the adiabatic index $\gamma = \text{const} > 1$.

The functions j, w and $\widehat{w}$ are given by the same formulas (16) but with new p whereas the regularized viscous stress Π and heat flux q are as follows:

$$j = \rho(u - w), \quad w = \frac{\tau}{\rho}\partial_x(\rho u^2 + p), \quad \widehat{w} = \frac{\tau}{\rho}(\rho u \partial_x u + \partial_x p), \tag{29}$$

$$\Pi = \mu\partial_x u + \rho u \widehat{w} + \tau(u\partial_x p + \gamma p \partial_x u), \tag{30}$$

$$-q = \widetilde{\varkappa}\partial_x \varepsilon + \tau\rho\Big(\partial_x\varepsilon - \frac{p}{\rho^2}\partial_x\rho\Big)u^2, \tag{31}$$

with the regularization parameter $\tau = \tau(\rho, u, \varepsilon) > 0$. The coefficients of artificial viscosity μ and (scaled) heat conductivity $\widetilde{\varkappa}$ are given by the standard QGD-formulas $\mu = \alpha_S\tau p$ and $\widetilde{\varkappa} = \widehat{\alpha}_P\gamma\tau p$, where $\widehat{\alpha}_P = \frac{1}{\alpha_P}$ and $\alpha_P > 0$ is the Prandtl number. Below the case $\widehat{\alpha}_P = 0$ is not excluded too.

The full QGD system is simplified to the compressible Navier-Stokes system, or the Euler one, for respectively $\tau = 0$ and $\mu > 0$, or $\tau = \mu = 0$.

The QGD system (27)–(31) can be linearized [14] at a constant solution $\rho_* > 0$, u_* and $\varepsilon_* > 0$ by setting $\rho = \rho_* + \rho_*\tilde{\rho}$, $u = u_* + \frac{c_*}{\sqrt{\gamma}}\tilde{u}$ and $\varepsilon = \varepsilon_* + \sqrt{\gamma - 1}\tilde{\varepsilon} > 0$, where $c_* = \sqrt{\gamma(\gamma - 1)\varepsilon_*}$ is the sound velocity, and taking $\tau_* = \tau(\rho_*, u_*, \varepsilon_*)$. The linearized system for the scaled small perturbations $\mathbf{z} = (\tilde{\rho}, \tilde{u}, \tilde{\varepsilon})^T$ can be written in the vector form (1) with the matrices

$$B = \begin{pmatrix} M & \frac{1}{\sqrt{\gamma}} & 0 \\ \frac{1}{\sqrt{\gamma}} & M & \sqrt{\frac{\gamma-1}{\gamma}} \\ 0 & \sqrt{\frac{\gamma-1}{\gamma}} & M \end{pmatrix}, \quad A = \begin{pmatrix} M^2 + \frac{1}{\gamma} & 2\frac{M}{\sqrt{\gamma}} & \frac{\sqrt{\gamma-1}}{\gamma} \\ 2\frac{M}{\sqrt{\gamma}} & M^2 + \widehat{\alpha}_S + 1 & 2\sqrt{\frac{\gamma-1}{\gamma}}M \\ \frac{\sqrt{\gamma-1}}{\gamma} & 2\sqrt{\frac{\gamma-1}{\gamma}}M & M^2 + \widehat{\alpha}_P + \frac{\gamma-1}{\gamma} \end{pmatrix}, \tag{32}$$

where $\widehat{\alpha}_S := \frac{\alpha_S}{\gamma}$. The result of the linearization of some schemes for the QGD system coincides with (3) with matrices (32), see [16, 18].

Recall the results on the necessary and sufficient conditions for L^2-dissipativity from [16] (in the modified form).

Theorem 5. *For scheme* (3) *with matrices* (32) *($n = 3$), the necessary condition* (7) *takes the form*

$$\tilde{\beta} \max\left\{2\alpha\underline{\lambda}_{\max}, \frac{1}{2\alpha}\right\} \leqslant 1, \tag{33}$$

and the sufficient condition (8) *is valid under the condition*

$$\tilde{\beta}\Big(2\alpha\overline{\lambda}_{\max} + \frac{1}{2\alpha}\Big) \leqslant 1, \tag{34}$$

for any $0 < \underline{\lambda}_{\max} \leqslant \lambda_{\max}(A) \leqslant \overline{\lambda}_{\max}$.

Here, in the particular case $u_* = 0$ (i.e., $M = 0$), one can take $\underline{\lambda}_{\max} = \lambda_{\max}(A) = \max\left\{\widehat{\alpha}_S + 1, \lambda(\widehat{\alpha}_P, \gamma)\right\} = \overline{\lambda}_{\max}$ *with*

$$\lambda(\widehat{\alpha}_P, \gamma) := \frac{\widehat{\alpha}_P + 1}{2} + \sqrt{\Big(\frac{\widehat{\alpha}_P - 1}{2}\Big)^2 + \frac{\gamma - 1}{\gamma}\widehat{\alpha}_P}.$$

In general, one can take

$$\underline{\lambda}_{\max} = M^2 + \max\Bigg\{\lambda(\widehat{\alpha}_P, \gamma),\ \frac{1}{2}\Big(\widehat{\alpha}_S + 1 + \frac{1}{\gamma}\Big) + \sqrt{\frac{1}{4}\Big(\widehat{\alpha}_S + 1 - \frac{1}{\gamma}\Big)^2 + \frac{4}{\gamma}M^2},$$
$$\frac{1}{2}\Big(\widehat{\alpha}_S - \frac{1}{\gamma} + \widehat{\alpha}_P\Big) + 1 + \sqrt{\frac{1}{4}\Big(\widehat{\alpha}_S + \frac{1}{\gamma} - \widehat{\alpha}_P\Big)^2 + 4\frac{\gamma - 1}{\gamma}M^2}\Bigg\}$$

and

$$\overline{\lambda}_{\max} = M^2 + \max\left\{\widehat{\alpha}_S + 1 + 2\left(\frac{1}{\sqrt{\gamma}} + \frac{\sqrt{\gamma - 1}}{\sqrt{\gamma}}\right)|M|, \widehat{\alpha}_P + 1 + 2\sqrt{\frac{\gamma - 1}{\gamma}}|M|\right\}.$$

Now we present a new enlarged sufficient condition for the validity of bound (4).

Theorem 6. *For scheme* (3) *with matrices* (32) *($n = 3$), bound* (4) *is valid under the condition*

$$\tilde{\beta}\left\{\alpha\overline{\lambda}_{\max} + \frac{1}{4\alpha} + \left[\Big(\alpha\overline{\lambda}_{\max} - \frac{1}{4\alpha}\Big)^2 + \frac{c_A^2}{4}\right]^{1/2}\right\} \leqslant 1 \tag{35}$$

for any $\lambda_{\max}(A) \leqslant \overline{\lambda}_{\max}$. *Here we have* $c_A > 0$ *and*

$$c_A^2 = \max\left\{\frac{a^2}{|A_{12}|}, \frac{b^2}{|A_{23}|}, \frac{\mathbf{b}^T A_{13}\mathbf{b}}{|A|}\right\}, \tag{36}$$

for $\widehat{\alpha}_P > 0$ *or* $M \neq 0$, *with the terms*

$$\begin{gathered}
a = -\tfrac{\widehat{\alpha}_S}{\sqrt{\gamma}}, \quad b = \big(\widehat{\alpha}_S - \widehat{\alpha}_P\big)\sqrt{\tfrac{\gamma-1}{\gamma}}, \\
\mathbf{b}^T A_{13}\mathbf{b} = \Big(\widehat{\alpha}_S - \widehat{\alpha}_P\sqrt{\tfrac{\gamma-1}{\gamma}}\,\Big)^2 M^2 + \Big(\widehat{\alpha}_S^2 + \tfrac{\gamma-1}{\gamma}\widehat{\alpha}_P\Big)\tfrac{\widehat{\alpha}_P}{\gamma} \geqslant 0, \\
|A_{12}| = M^4 + \Big(\widehat{\alpha}_S + 1 - \tfrac{3}{\gamma}\Big)M^2 + \tfrac{1}{\gamma}(\widehat{\alpha}_S + 1) > 0, \\
|A_{23}| = M^4 + \Big(\widehat{\alpha}_S + \widehat{\alpha}_P + 1 - 3\tfrac{\gamma-1}{\gamma}\Big)M^2 + (\widehat{\alpha}_S + 1)\Big(\widehat{\alpha}_P + \tfrac{\gamma-1}{\gamma}\Big) > 0, \\
|A| = M^6 + (\widehat{\alpha}_S + \widehat{\alpha}_P - 2)M^4 + \Big[\Big(\widehat{\alpha}_S + 1 - \tfrac{3}{\gamma}\Big)\widehat{\alpha}_P + \widehat{\alpha}_S + 1\Big]M^2 + (\widehat{\alpha}_S + 1)\tfrac{\widehat{\alpha}_P}{\gamma} \geqslant 0,
\end{gathered} \tag{37}$$

moreover, $|A| > 0$ *for* $\widehat{\alpha}_P > 0$ *or* $M \neq 0$.

For $\widehat{\alpha}_P = M = 0$, *one has* $\mathbf{b}^T A_{13}\mathbf{b} = |A| = 0$ *and the third term in the definition* (36) *of* c_A^2 *should be omitted.*

Proof. It is known [16] and straightforward to check that

$$A = B^2 + D, \quad D = \operatorname{diag}\big\{0, \widehat{\alpha}_S, \widehat{\alpha}_P\big\} \geqslant 0.$$

Moreover, $A > 0$ for $\widehat{\alpha}_P > 0$ or $M \neq 0$ (otherwise $A \geqslant 0$ and $|A| = 0$). Therefore it easy to calculate that

$$[A, B] = [D, B] = \begin{pmatrix} 0 & a & 0 \\ -a & 0 & b \\ 0 & -b & 0 \end{pmatrix}$$

with a and b given in (37).

Owing to inequality (12) and formulas (13) (see also Remark 3 for $\widehat{\alpha}_P = M = 0$ when $|A| = 0$), we have $\pm\mathbf{i}[A, B] \leqslant c_A A$ with $c_A \geqslant 0$ presented in the statement of the theorem (note that $\mathbf{b}^T A_{13}\mathbf{b}$ is given in the reduced form). Then Theorem 2 together with non-decreasing of the function $\lambda + a_0 + \sqrt{(\lambda - a_0)^2 + c^2}$ in λ imply the result.

Corollary 1. *The sufficient condition* (35) *is wider than* (34).

Proof. It is not difficult to check the lower bounds

$$\begin{gathered}
|A_{12}| > \frac{\widehat{\alpha}_S}{\gamma}, \quad |A_{23}| > (\widehat{\alpha}_S + \widehat{\alpha}_P)\frac{\gamma - 1}{\gamma}, \\
\underline{\lambda}_{\max} \geqslant M^2 + \max\left\{\widehat{\alpha}_S + 1, \widehat{\alpha}_P + \frac{\gamma - 1}{\gamma}\right\} > \max\left\{\widehat{\alpha}_S, \widehat{\alpha}_P\right\}.
\end{gathered}$$

Thus

$$a^2 < |A_{12}|\underline{\lambda}_{\max}, \quad b^2 < |A_{23}|\underline{\lambda}_{\max}.$$

Moreover, we have

$$\mathbf{b}^T A_{13}\mathbf{b} \leqslant \max\left\{\widehat{\alpha}_S^2, \frac{\gamma-1}{\gamma}\widehat{\alpha}_P^2\right\}M^2 + \max\left\{\widehat{\alpha}_S, \widehat{\alpha}_P\right\}\left(\widehat{\alpha}_S + \frac{\gamma-1}{\gamma}\right)\frac{\widehat{\alpha}_P}{\gamma}. \tag{38}$$

Next using the inequality $M^4 + \frac{1}{\gamma^2} \geqslant \frac{2}{\gamma}M^2$, we get one more lower bound

$$\begin{aligned}|A| &\geqslant M^2(M^2-1)^2 + \widehat{\alpha}_S(\widehat{\alpha}_P+1)M^2 + \widehat{\alpha}_P\left[M^4 + \left(1-\frac{3}{\gamma}\right)M^2 + \frac{1}{\gamma}\right] + \widehat{\alpha}_S\frac{\widehat{\alpha}_P}{\gamma}\\ &\geqslant \left[\widehat{\alpha}_S(\widehat{\alpha}_P+1) + \widehat{\alpha}_P\frac{\gamma-1}{\gamma}\right]M^2 + \left(\widehat{\alpha}_S + \frac{\gamma-1}{\gamma}\right)\frac{\widehat{\alpha}_P}{\gamma}.\end{aligned}$$

Consequently for $\widehat{\alpha}_P > 0$ or $M \neq 0$

$$|A|\underline{\lambda}_{\max} \geqslant \left[\widehat{\alpha}_S^2(\widehat{\alpha}_P+1) + \widehat{\alpha}_P^2\frac{\gamma-1}{\gamma}\right]M^2 + \max\left\{\widehat{\alpha}_S, \widehat{\alpha}_P\right\}\left(\widehat{\alpha}_S + \frac{\gamma-1}{\gamma}\right)\frac{\widehat{\alpha}_P}{\gamma},$$

and from (38) we see that

$$\mathbf{b}^T A_{13}\mathbf{b} \leqslant |A|\underline{\lambda}_{\max}.$$

One can check that the inequality is strict for $\widehat{\alpha}_P > 0$ or $M \neq 0$.

Thus $c_A < \underline{\lambda}_{\max}$ and owing to Remark 1 the sufficient condition (35) is broader than (34).

Notice also that once again $c_A = O\left(\frac{1}{M^2}\right)$ as $M \to \infty$ and, moreover, the sufficient condition (35) rapidly tends to the necessary condition (33) (provided that $\underline{\lambda}_{\max} = \lambda_{\max}(A) = \overline{\lambda}_{\max}$ are taken) as $|M|$ grows.

In terms of β and $\widehat{\alpha}$, see formulas (22), the sufficient condition (21) is rewritten in the form

$$\beta\left\{\widehat{\alpha}\frac{\overline{\lambda}_{\max}}{(|M|+1)^2} + \frac{1}{4\widehat{\alpha}} + \left[\left(\widehat{\alpha}\frac{\overline{\lambda}_{\max}}{(|M|+1)^2} - \frac{1}{4\widehat{\alpha}}\right)^2 + \frac{c_A^2}{4(|M|+1)^2}\right]^{1/2}\right\} \leqslant 1.$$

The optimal value $\widehat{\alpha} = \widehat{\alpha}_{opt} \equiv \frac{|M|+1}{2\sqrt{\overline{\lambda}_{\max}(A)}}$ leads to the minimal value $\frac{\sqrt{\overline{\lambda}_{\max}(A)}+\frac{1}{2}c_A}{|M|+1}$ of the expression in the curly brackets.

For the above alternative choice $\tau_* = \frac{a_0}{2}\Delta t$ (independent of h), the necessary condition (33) is transformed into (24) with substituting $\underline{\lambda}_{\max}$ for $\lambda_{\max}(A)$, and the sufficient conditions (34) and (35) are transformed into (25) and (26), respectively, with substituting $\overline{\lambda}_{\max}$ for $\lambda_{\max}(A)$.

Acknowledgements The study was supported by the Russian Foundation for Basic Research, project nos. 19-11-00169 and 18-01-00587.

References

1. Abgrall, R., Shu, C.-W. (eds.): Handbook of Numerical Methods for Hyperbolic Problems: Basic And Fundamental Issues. North Holland, Amsterdam (2016)
2. Chetverushkin, B.N.: Kinetic Schemes and Quasi-Gas Dynamic System of Equations. CIMNE, Barcelona (2008)
3. Godunov, S.K., Ryabenkii, V.S.: Difference Schemes. Studies in Mathematics and its Applications **19**, North Holland, Amsterdam (1987)
4. Guermond, J.-L., Popov, B.: Viscous regularization of the Euler equations and entropy principles. SIAM J. Appl. Math. **74**(2), 284–305 (2014)
5. Elizarova, T.G.: Quasi-Gas Dynamic Equations. Springer, Dordrecht (2009)
6. Kulikovskii, A.G., Pogorelov, N.V., Semenov, A.Yu.: Mathematical Aspects of Numerical Solution of Hyperbolic Systems. Chapman and Hall/CRC, London (2001)
7. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge University Press, Cambridge (2004)
8. Richtmyer, R.D., Morton, K.W.: Difference Methods For Initial-Value Problems, 2nd edn. Wiley-Interscience, New York (1967)
9. Sheretov, Yu.V.: Continuum Dynamics Under Spatiotemporal Averaging. RKhD, Moscow-Izhevsk (2009). [in Russian]
10. Suhomozgii, A.A., Sheretov, Yu.V.: Stability analysis of a finite-difference scheme for solving the Saint-Venant equations in the shallow water theory. In: Applications of Functional Analysis in Approximation Theory, Tver State Univ. (2013) 48–60 [in Russian]
11. Svärd, M.: A new Eulerian model for viscous and heat conducting compressible flows. Phys. A. Stat. Mech. Appl. **506**, 350–375 (2018)
12. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics, 3rd edn. Springer, Berlin (2009)
13. Zlotnik, A.A.: Energy equalities and estimates for barotropic quasi-gasdynamic and quasi-hydrodynamic systems of equations. Comput. Math. Math. Phys. **50**(2), 310–321 (2010)
14. Zlotnik, A.A., Chetverushkin, B.N.: Parabolicity of the quasi-gasdynamic system of equations, its hyperbolic second-order modification, and the stability of small perturbations for them. Comput. Math. Math. Phys. **48**(3), 420–446 (2008)
15. Zlotnik, A., Lomonosov, T.: On conditions for weak conservativeness of regularized explicit finite-difference schemes for 1D barotropic gas dynamics equations. In: Pinelas, S., et al. (eds.) Differential and Difference Equations with Applications, Springer Proceedings in Mathematics & Statistics **230** 635–647 (2018)
16. Zlotnik, A.A., Lomonosov, T.A.: On conditions for L^2-dissipativity of linearized explicit QGD finite-difference schemes for one-dimensional gas dynamics equations. Dokl. Math. **98**(2), 458–463 (2018)
17. Zlotnik, A.A., Lomonosov, T.A.: Conditions for L^2-dissipativity of linearized explicit difference schemes with regularization for 1D barotropic gas dynamics equations. Comput. Math. Math. Phys. **59**(3), 452–464 (2019)
18. Zlotnik, A.A., Lomonosov, T.A.: Verification of an entropy dissipative QGD-scheme for the 1D gas dynamics equations. Math. Model. Anal. **24**(2), 179–194 (2019)

Equilibrium of a Linearly Elastic Body Under Generalized Boundary Data

Giulio Starita and Alfonsina Tartaglione

Abstract We consider the interior Dirichlet, Neumann and Robin problems associated to the differential system of linear elastostatics with singular data. We prove that if the assigned displacement field a on the C^2 boundary S of the reference configuration of the elastic body belongs to $W^{-1/2,2}(S)$, then there exists a unique solution to the equilibrium problem which takes the boundary datum a in a well–defined sense; similar results hold if we assign the traction or a linear combination of displacement and traction on the boundary. Moreover, natural estimates controlling the norms of the solutions with the norms of the data hold and analogous results are obtained for the exterior problems requiring the displacement vanishes at infinity.

Keywords Linear elastostatics · Boundary value problems · Singular data · Layer potentials

1 Introduction

Let S be a closed C^2–surface in $\mathbb{R}^3$. It splits the space into an interior domain Ω_i, which is bounded, and an unbounded exterior domain Ω_e. Let Ω_i or Ω_e represent the reference configuration of a linearly elastic body whose material properties are collected in the (constant) components C_{ijhk} of the *elasticity tensor* $\mathbb{C}$ associated to the body. We will refer to Ω when we are considering Ω_i or Ω_e, indifferently.

The equilibrium configurations of the body are represented by the domains $\{x + u(x),\ x \in \Omega\}$ with u solutions to the system of partial differential equations [4]

$$\operatorname{div} \mathbb{C}(\nabla u) = 0, \quad \text{in } \Omega, \tag{1}$$

G. Starita · A. Tartaglione (✉)
Dipartimento di Matematica e Fisica, Università degli Studi della Campania “Luigi Vanvitelli”, Caserta, Italy
e-mail: alfonsina.tartaglione@unicampania.it

G. Starita
e-mail: giulio.starita@unicampania.it

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_6

which in components writes

$$\partial_j(C_{ijhk}\partial_k u_h) = 0, \quad i = 1, 2, 3, \tag{2}$$

where summation over repeated indexes is understood. Observe that, for simplicity, we are supposing that no body forces act on Ω.

The elasticity tensor $\mathbb{C}$ is *positive definite* if[1]

$$E \cdot \mathbb{C}(E) \geq |\mathrm{sym}E|^2, \quad \forall\, E \in \mathrm{Lin}, \tag{3}$$

and *strongly elliptic* if

$$(a \otimes b) \cdot \mathbb{C}(a \otimes b) > 0, \quad \forall\, a, b \neq 0. \tag{4}$$

Unless otherwise specified we suppose that $\mathbb{C}$ is strongly elliptic.

To determine the solutions of system (1) we have to associate to (1) appropriate boundary conditions, we consider of the following classical type:

$$u = a \quad \text{on } S, \tag{5}$$

or

$$\mathbb{C}(\nabla u)n = s \quad \text{on } S, \tag{6}$$

where n is the unit vector, normal to S and pointing out of Ω_i (so that, n is the exterior normal for $\Omega = \Omega_i$ and the interior normal for $\Omega = \Omega_e$). If $\Omega = \Omega_e$, besides a boundary condition, we have to require a suitable condition at infinity. We suppose

$$\lim_{|x| \to +\infty} u(x) = 0. \tag{7}$$

Problem (1)–(5) is known as *Dirichlet problem* (or *displacement problem*) of linear elastostatics; problem (1)–(6) is known as *Neumann problem* (or *traction problem*) of linear elastostatics. We talk about an *interior problem* when $\Omega = \Omega_i$ and *exterior problem* when $\Omega = \Omega_e$. In this last case we always require that (7) is met (cf. [9] and [10]).

When the boundary data belong to the trace spaces associated to the Sobolev spaces where we are looking for the solutions, the existence and uniqueness results for the interior and the exterior Dirichlet or Neumann boundary value problems are well–established.

[1] We will use standard notation as in [4]. Moreover, $W^{k,q}(\Omega)$ is the Sobolev space of all $\varphi \in L^1_{\mathrm{loc}}(\Omega)$ such that $\|\varphi\|_{W^{k,q}(\Omega)} = \|\varphi\|_{L^q(\Omega)} + \|\nabla_k\varphi\|_{L^q(\Omega)} < +\infty$; $W_0^{k,q}(\Omega)$ is the completion of $C_0^\infty(\Omega)$ with respect to $\|\varphi\|_{W^{k,q}(\Omega)}$ and $W^{-k,q'}(\Omega)$, with $1/q + 1/q' = 1$, is its dual space. $W^{k-1/q,q}(S)$ is the trace space of $W^{k,q}(\Omega)$ and $W^{1-k-1/q',q'}(S)$ is its dual space.

In particular, referring to a solution to (1) as a *weak solution* to (1), *i.e.* a field $u \in W^{1,2}_{loc}(\Omega)$ satisfying

$$\int_{\Omega} \nabla u \cdot \mathbb{C}(\nabla\varphi) = 0, \quad \forall\varphi \in C_0^{\infty}(\Omega), \tag{8}$$

the variational theory leads to the following theorems for the interior problems with regular data (see [1–3, 7, 8, 11]).

Theorem 1. *If $a \in W^{3/2,2}(S)$, then the displacement problem*

$$\begin{aligned} \operatorname{div} \mathbb{C}[\nabla u] = 0 \quad &\text{in } \Omega_i, \\ u = a \quad &\text{on } S, \end{aligned} \tag{9}$$

has a unique solution $u \in W^{2,2}(\Omega_i)$ and

$$\|u\|_{W^{2,2}(\Omega_i)} \le c\|a\|_{W^{3/2,2}(S)}. \tag{10}$$

Theorem 2. *If $\mathbb{C}$ is positive definite and $s \in W^{1/2,2}(S)$ satisfies*

$$\int_S \rho \cdot s = 0 \tag{11}$$

for all ρ infinitesimal rigid displacements, then the traction problem

$$\begin{aligned} \operatorname{div} \mathbb{C}[\nabla u] = 0 \quad &\text{in } \Omega_i, \\ \mathbb{C}(\nabla u)n = s \quad &\text{on } S, \end{aligned} \tag{12}$$

has a unique solution[2] *$u \in W^{2,2}(\Omega_i)$ and*

$$\|u\|_{W^{2,2}(\Omega_i)} \le c\|s\|_{W^{1/2,2}(S)} \tag{13}$$

In many situations, surely more interesting in the applications, the boundary data are nevertheless represented by non–regular fields, as, for example, in the case of concentrated loads. The aim of the present paper is just to illustrate how the various problems can be treated in the presence of singular data. In particular, we will present existence and uniqueness results via the methods of the potential theory. These ones will be applied to layer potentials suitably defined on singular densities and will be mixed to the existence and uniqueness theorems for regular data mentioned above in order to obtain our results for generalized data.

[2] Observe that for solutions to the traction problem we mean “normalized” displacements (see [4]).

2 Preliminaries

In this section we recall the definitions of the elastic layer potentials and some properties we will need in the sequel.

Since the elasticities are supposed to be constant, system (1) admits a fundamental solution [5] $U(x-y)$, which is a regular solution for $x \neq y$ to

$$\operatorname{div} \mathbb{C}(\nabla U(x-y)) = \delta(x-y)$$

with δ the Dirac function. The asymptotic behaviour of U is clarified by its expression for isotropic bodies.[3] Indeed, in this case,

$$U(x-y) = \frac{1}{16\pi(1-\nu)|x-y|}\left\{(3-4\nu)\mathbf{1} + \frac{(x-y)\otimes(x-y)}{|x-y|^2}\right\}, \tag{14}$$

where $\nu = \lambda/2(\lambda+\mu)$ is the *Poisson ratio*.

The *single layer potential* of density $\psi \in L^1(S)$ is defined as the field

$$v[\psi](x) = \int_S U(x-\zeta)\psi(\zeta)\mathrm{d}\sigma_\zeta, \tag{15}$$

and the *double layer potential* of density $\varphi \in L^1(S)$ as

$$w[\varphi](x) = \int_S \mathbb{C}(\nabla U(x-\zeta))(\varphi\otimes n)(\zeta)\mathrm{d}\sigma_\zeta, \tag{16}$$

The single layer potential is an analytical solution of (1) in $\mathbb{R}^3\backslash S$ and its "interior limit" is equal to the "exterior limit":

$$\lim_{\epsilon\to 0^+} v[\psi](\xi - \epsilon l(\xi)) = \lim_{\epsilon\to 0^+} v[\psi](\xi + \epsilon l(\xi)) = \mathscr{S}[\psi](\xi), \quad \text{a.e. on } S \tag{17}$$

where l is a generic axis in a ball tangent (on the side of the normal vector n) to S at ξ.

Relation (17) shows that $v[\psi]$ is continuous in $\mathbb{R}^3$ and, since

$$\|v[\psi]\|_{W^{2,2}(\Omega_i)} \leq c\|\psi\|_{W^{1/2,2}(S)}, \tag{18}$$

for some constants c depending only on Ω, the map

$$\mathscr{S} : W^{1/2,2}(S) \to W^{3/2,2}(S) \tag{19}$$

[3] For isotropic bodies, $\mathbb{C}(E) = 2\mu\,\mathrm{sym}E + \lambda(\mathrm{tr}E)\mathbf{1}$, $\forall\, E \in \mathrm{Lin}$. λ and μ are called the *Lamé moduli*.

with $\mathscr{S}[\psi](\xi)$ equal to the limit (17) is a well–defined linear and continuous operator. Thinking of its meaning, we refer to $\mathscr{S}$ as the *trace operator* associated with the single layer potential.

Analogously, the double layer potential is an analytical solution of (1) in $\mathbb{R}^3\backslash S$ and the "interior limit" of the associated traction is equal to the "exterior limit":

$$\lim_{\epsilon\to 0^+} \mathbb{C}(\nabla w)(\xi - \epsilon l(\xi))n(\xi) = \lim_{\epsilon\to 0^+} \mathbb{C}(\nabla w)(\xi + \epsilon l(\xi))n(\xi) = \mathscr{Z}[\varphi](\xi), \tag{20}$$

a.e. on S, so that the traction of the double layer potential is continuous in $\mathbb{R}^3$ and, since

$$\|w[\varphi]\|_{W^{2,2}(\Omega)} \leq c\|\varphi\|_{W^{1/2,2}(\partial\Omega)}, \tag{21}$$

for some constants c depending only on Ω, the map

$$\mathscr{Z} : W^{1/2,2}(S) \to W^{1/2,2}(S) \tag{22}$$

with $\mathscr{Z}[\varphi](\xi)$ equal to the limit (20) is a well–defined linear and continuous operator we refer to as the *trace operator* associated with the traction of the double layer potential.

In contrast, the double layer potential and the traction of the single layer potential present jumps across the surface S. Precisely, the following limits exist

$$\lim_{\epsilon\to 0^+} w[\varphi](\xi \mp \epsilon l(\xi)) = \mathscr{W}^{\pm}[\psi](\xi), \tag{23}$$

$$\lim_{\epsilon\to 0^+} \mathbb{C}(\nabla v[\psi])(\xi \mp \epsilon l(\xi))n(\xi) = \mathscr{T}^{\pm}[\psi](\xi) \tag{24}$$

a.e. on S, define the linear and continuous operators

$$\mathscr{W}^{\pm} : W^{3/2,2}(S) \to W^{3/2,2}(S) \tag{25}$$

$$\mathscr{T}^{\pm} : W^{1/2,2}(S) \to W^{1/2,2}(S) \tag{26}$$

and the following jump conditions hold

$$\psi = \mathscr{T}^{+}[\psi] - \mathscr{T}^{-}[\psi], \tag{27}$$

$$\varphi = \mathscr{W}^{+}[\varphi] - \mathscr{W}^{-}[\varphi]. \tag{28}$$

By applying the methods of the potential theory in place of those ones of the variational theory leading to Theorems 1 and 2, we can look at these theorems as representation results for the solutions to the boundary value problems when the data are regular. Indeed, we can look for the solutions to the Dirichlet problem or to the Neumann problem in terms of single or double layer potentials. Observe that the requirement (11) in order to obtain existence and uniqueness for the interior Neumann problem corresponds to the orthogonality condition with the elements of the kernel of $\mathscr{Z}$ imposed to the boundary datum.[4] This observation suggests us to consider the extensions of the maps $\mathscr{S}$ and $\mathscr{Z}$ to spaces of singular densities and to verify the possibility to apply the methods of the potential theory (in particular, the Fredholm alternative) to get existence and uniqueness of solutions to the boundary value problems in the case of generalized boundary data.

3 The Dirichlet Problem

In virtue of the continuity of $\mathscr{S}$,

$$\|\mathscr{S}[\psi]\|_{W^{3/2,2}(S)} \le c\|\psi\|_{W^{1/2,2}(S)} \tag{29}$$

with $c = c(\Omega)$. Let $\psi \in W^{-3/2,2}(S)$ and ψ_k a regular sequence which converges to ψ strongly in $W^{-3/2,2}(S)$. By (29)

$$\left|\int_S \phi \cdot \mathscr{S}[\psi_k]\right| = \left|\int_S \psi_k \cdot \mathscr{S}[\phi]\right| \le c\|\psi_k\|_{W^{-3/2,2}(S)}\|\phi\|_{W^{1/2,2,}(S)}.$$

Therefore, $\mathscr{S}$ can be extended to a linear and continuous operator

$$\mathscr{S}^* : W^{-3/2,2}(S) \to W^{-1/2,2}(S),$$

which coincides with the adjoint of $\mathscr{S}$. We can think of $\mathscr{S}^*$ as the trace operator associated with the *extended* single layer potential with density $\psi \in W^{-3/2,2}(S)$:

$$v^*[\psi](x) = < U(x,\cdot), \psi > \tag{30}$$

where $<,>$ denotes the duality pairing between $W^{3/2,2}(S)$ and its dual space $W^{-3/2,2}(S)$. Observe that by (18) it follows that

$$\|v[\psi]\|_{L^2(\Omega)} \le c\|\psi\|_{W^{-3/2,2}(S)}. \tag{31}$$

If we are able to prove that $\mathscr{S}^*$ is Fredholmian[5] we can then obtain an existence and uniqueness result for the Dirichlet problem with datum in the range of $\mathscr{S}^*$ by looking for a solution of the form (30), then translating the problem into the functional equation

$$\mathscr{S}^*[\psi] = a \tag{32}$$

[4] The boundary datum a for the interior Dirichlet problem is not required to satisfy any compatibility conditions since the kernel of $\mathscr{S}$ is reduced to the null vector field.

[5] This means that it has closed range and zero index.

and finally applying to (32) the Fredholm alternative. So that our result will be a consequence of the following key lemma (see [12]).

Lemma 1. *The operator $\mathscr{S}$ is Fredholmian and* $\ker\mathscr{S} = \ker\mathscr{S}^* = \{0\}$.

With the results of Lemma 1 at hand, we obtain the following existence and uniqueness result [15] (see also [14, 16]).

Theorem 3. *If $a \in W^{-1/2,2}(S)$ then the Dirichlet problem* (1)–(5) *has a solution u expressed by a simple layer potential with density $\psi \in W^{-3/2,2}(S)$, attaining the boundary datum in the sense of* (32). *It satisfies the estimate*

$$\|u\|_{L^2(\Omega)} \leq c\|a\|_{W^{-1/2,2}(S)} \tag{33}$$

and is unique in the class of all $u \in L^2(\Omega)$ such that[6]

$$\int_\Omega u \cdot \phi = \pm\langle a, \mathbb{C}[\nabla z]n\ \rangle, \tag{34}$$

for all $\phi \in C_0^\infty(\Omega)$, with z solution of

$$\begin{aligned} \operatorname{div}\mathbb{C}[\nabla z] = \phi \quad &\text{in } \Omega, \\ z = 0 \quad &\text{on } \partial\Omega, \end{aligned} \tag{35}$$

and $z = o(1)$ if $\Omega = \Omega_e$.

By observing that the adjoint operators of $\mathscr{T}^\pm$ in (26) are the operators

$$\mathscr{W}^{*\mp} : W^{-1/2,2}(S) \to W^{-1/2,2}(S),$$

which represent the traces of the double layer potential with density in $W^{-1/2,2}(S)$, we can also look for the solution of the Dirichlet problem in term of a double layer potential. In this case we prove the existence and uniqueness by analysing the functional equations

$$\mathscr{W}^{*+}[\psi] = a$$

for $\Omega = \Omega_i$ and

$$\mathscr{W}^{*-}[\psi] = a$$

for $\Omega = \Omega_e$ (see [12]).

[6] In (34) $+$ is for $\Omega = \Omega_i$ and $-$ for $\Omega = \Omega_e$.

4 The Neumann Problem

Reasoning as we did in Sect. 3, we obtain an existence and uniqueness result with singular datum also for the Neumann problem (1)–(6).

First of all we consider the extension of the operator $\mathscr{L}$

$$\mathscr{L}^* : W^{-1/2,2}(S) \to W^{-1/2,2}(S)$$

which coincides with its adjoint operator. It defines the trace of the traction field of the *extended* double layer potential with density $\varphi \in W^{-1/2,2}(S)$:

$$w^*[\varphi](x) = < \mathbb{C}(\nabla U(x,\cdot))n, \varphi > \tag{36}$$

with $<,>$ the duality pairing between $W^{1/2,2}(S)$ and $W^{-1/2,2}(S)$. Then we prove that $\mathscr{L}^*$ is Fredholmian. Precisely, the following lemma holds (see [12]).

Lemma 2. *The operator $\mathscr{L}^*$ is Fredholmian and*

$$\ker\mathscr{L} = \ker\mathscr{L}^* = \mathfrak{R}. \tag{37}$$

where $\mathfrak{R}$ is the set of the infinitesimal rigid displacements.

In virtue of Lemma 2 we can look for a solution in form of a double layer potential $u = w^*[\varphi]$ and apply the Fredholm alternative to the functional equation

$$\mathscr{L}^*[\varphi] = s. \tag{38}$$

We obtain the following existence and uniqueness result [13].

Theorem 4. *Let $\mathbb{C}$ be positive definite. If $s \in W^{-1/2,2}(S)$ and*

$$< s, \rho > = 0, \quad \forall\, \rho \in \mathfrak{R}, \tag{39}$$

for $\Omega = \Omega_i$, then (1)–(6) *has a solution expressed by a double layer potential with density $\psi \in W^{-1/2,2}(S)$, attaining the boundary datum in the sense of* (38)*. It satisfies the estimate*

$$\|u\|_{L^2(\Omega)} \le c\|s\|_{W^{-1/2,2}(S)} \tag{40}$$

and is unique in the class of all $u \in L^2_{\rm loc}(\Omega)$ such that[7]

$$\int_\Omega u \cdot \phi = \mp\!< s, z >, \tag{41}$$

[7] In (45) $-$ is for $\Omega = \Omega_i$ and $+$ for $\Omega = \Omega_e$.

for all $\phi \in C_0^\infty(\Omega)$, with z solution of

$$\begin{aligned} \operatorname{div} \mathbb{C}(\nabla z) &= \phi \quad \text{in } \Omega, \\ \mathbb{C}(\nabla z)n &= 0 \quad \text{on } S \end{aligned} \tag{42}$$

and $z = o(1)$ if $\Omega = \Omega_e$.

Observe that we can also look for the solution in form of a single layer potential. In this case we take into account that the adjoint operators of $\mathscr{W}^{\pm}$ in (25) are the operators

$$\mathscr{T}^{*\mp} : W^{-3/2,2}(S) \to W^{-3/2,2}(S),$$

which represent the traces of the traction of the simple layer potential with density in $W^{-3/2,2}(S)$. The existence result then follows from the analysis of the functional equation

$$\mathscr{T}^{*+}[\psi] = s$$

for $\Omega = \Omega_i$ and

$$\mathscr{T}^{*-}[\psi] = s$$

for $\Omega = \Omega_e$ (see [12]).

5 The Robin Problem

Boundary conditions other than the classical (5) and (6) can be considered. For instance, we can analyse the Robin problem, which consists in finding a solution to (1) such that a linear combination of displacement and traction is assigned on the boundary:

$$\alpha u + \mathbb{C}(\nabla u)n = t \quad \text{on } S \tag{43}$$

with $\alpha > 0$. To prove the existence and uniqueness in the case, say, $t \in W^{-3/2,2}(S)$, we look for a solution in form of a single layer potential $u = v^*[\psi]$. Condition (43) then writes

$$(\alpha \mathscr{S}^* + \mathscr{T}^{*+})[\psi]$$

for $\Omega = \Omega_i$. Since $\mathscr{S}^*$ is compact from $W^{-3/2,2}(S)$ into itself, $\alpha\mathscr{S}^* + \mathscr{T}^{*+}$ is a compact perturbation of a Fredholmian operator. So it is Fredholmian and the existence then follows from the analysis of the kernel of the adjoint operator $\alpha\mathscr{S} - \mathscr{W}^-$. Precisely, the following theorem holds (see [13]).

Theorem 5. *Let $\mathbb{C}$ be positive definite. If $t \in W^{-3/2,2}(S)$, then (1)–(43) has a solution expressed by a double layer potential with density $\psi \in W^{-3/2,2}(S)$. It satisfies the estimate*

$$\|u\|_{L^2(\Omega)} \le c\|t\|_{W^{-3/2,2}(S)} \tag{44}$$

and is unique in the class of all $u \in L^2_{\text{loc}}(\Omega)$ *such that*[8]

$$\int_\Omega u \cdot \phi = \mp\langle t, z\rangle, \tag{45}$$

for all $\phi \in C_0^\infty(\Omega)$*, with* z *solution of*

$$\begin{aligned} \operatorname{div} \mathbb{C}(\nabla z) &= \phi \quad \text{in } \Omega, \\ \alpha z + \mathbb{C}(\nabla z)n &= 0 \quad \text{on } S \end{aligned} \tag{46}$$

and $z = o(1)$ *if* $\Omega = \Omega_e$.

Remark 1. If the boundary of the elastic body is more regular, that is if S is a C^k–surface with $k > 2$, then with the methods used in Theorems 3, 4, 5, existence and uniqueness is achieved for boundary data with more severe singularities, *i.e.* belonging to the general spaces $W^{1-k-1/q,q}(S),\ q \in (1, +\infty)$ [13].

Remark 2. If the boundary of the elastic body is less regular, then we obtain existence and uniqueness results in the case of isotropic bodies. In particular, if S is a $C^{1,\alpha}$–surface, then by applying classical results by Kupradze and Mikhlin [6], we can prove that there exist unique solutions to the boundary value problems with data in $L^q(S)$. For the Neumann problem we can assume $s \in W^{-1,q}(S)$ [13].

Remark 3. If the body is isotropic and the boundary S is a C^1–surface, we can prove the existence of a solution to the displacement problem in form of a double layer potential defined through the *pseudostress* field [13].

6 Conclusions

In this paper we presented existence and uniqueness results on the solutions to the Dirichlet, the Neumann and the Robin boundary value problems associated to the differential system of elastostatics. The novelty is that the theorems are obtained in the presence of generalized data, which better represent the data deriving from the applications related to the equilibrium of an elastic body. The analysis of the boundary value problems is done by investigating the functional equations representing the attainability of the boundary data and examinating the kernels of the involved operators.

[8] In (45) $-$ is for $\Omega = \Omega_i$ and $+$ for $\Omega = \Omega_e$.

References

1. Fichera, G.: Sull'esistenza e sul calcolo delle soluzioni dei problemi al contorno, relativi all'equilibrio di un corpo elastico. Ann. Sc. Norm. Super. Pisa Cl. Sci. **III**(4), 35–99 (1950)
2. Fichera, G.: Existence theorems in elasticity. In: Truesedell. C. (ed.) Handbuch der Physik, vol. VIa/2. Springer (1972)
3. Giusti, E.: Metodi diretti nel calcolo delle variazioni, Unione Matematica Italiana (1994). English translation - Direct methods in the calculus of variations. Word Scientific (2004)
4. Gurtin, M.E.: The linear theory of elasticity. In: Truesedell, C. (ed.) Handbuch der Physik, vol. VIa/2. Springer (1972)
5. John, F.: Plane waves and spherical means applied to partial differential equations. Interscience (1955)
6. Kupradze, V.D., Gegelia, T.G., Basheleishvili, M.O., Burchuladze, T.V.: Three Dimensional Problems of the Mathematical Theory of Elasticity and Thermoelasticity. North-Holland (1979)
7. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary-value Problems and Applications, vol. I. Springer, Cham (1972)
8. Nečas, J.: Les méthodes directes en théorie des équations élliptiques. Masson-Paris and Academie-Prague (1967)
9. Russo, A., Tartaglione, A.: On the contact problem of classical elasticity. J. Elasticity **99**, 19–38 (2010)
10. Russo, A., Tartaglione, A.: Strong uniqueness theorems and the Phragmen-Lindelof principle in nonhomogeneous elastostatics. J. Elasticity **102**, 133–149 (2011)
11. Russo, A., Tartaglione, A.: On the Stokes problem with data in L^1. Zeitschrift fur Angewandte Mathematik und Physik **64**(4), 1327–1336 (2013)
12. Starita, G., Tartaglione, A.: On the Fredholm property of the trace operators associated with the elastic layer potentials. Mathematics **7**, 134 (2019)
13. Starita, G., Tartaglione, A.: Boundary value problems in elastostatics with singular data. Lith. Math. J. **60**(3), 396–409 (2020). https://doi.org/10.1007/s10986-020-09489-3
14. Tartaglione, A.: On the Stokes and Oseen problems with singular data. J. Math. Fluid Mech. **16**, 407–417 (2014)
15. Tartaglione, A.: A note on the displacement problem of elastostatics with singular boundary values. Axioms **8**, 46 (2019)
16. Russo, A., Tartaglione, A.: On the existence of singular solutions of the stationary Navier-Stokes problem. Lith. Math. J. **53**(4), 423–437 (2013)

Nonlocal Problems for the Fourth Order Impulsive Partial Differential Equations

Anar T. Assanova, Aziza D. Abildayeva, and Agila B. Tleulessova

Abstract Nonlocal problems for an impulsive system of fourth-order partial differential equations are investigated. By the method of introducing additional functions, the problems under study are reduced to an equivalent problem consisting of the impulsive system of second-order hyperbolic equations and integral relations. Algorithm for finding the approximate solutions to the equivalent problem is constructed and its convergence is proved. Sufficient conditions are obtained for the unique solvability of a nonlocal problem for the impulsive system of fourth-order partial differential equations. As an example, the conditions for the unique solvability of a periodic problem for the impulsive system of fourth-order partial differential equations are established.

Keywords Impulsive partial differential equations · Nonlocal problems · Periodic problem

1 Introduction

As is well-known, various problems of the dynamics and kinetics of gas sorption, drying processes by air stream, the movement of adsorbed mixtures, and etc., lead to boundary value problems for impulsive systems of differential equations. Periodic

A. T. Assanova (✉) · A. D. Abildayeva · A. B. Tleulessova
Institute of Mathematics and Mathematical Modeling,
125 Pushkin str., A26G7T5 Almaty, Kazakhstan
e-mail: anartasan@gmail.com

A. D. Abildayeva
e-mail: azizakz@mail.ru

A. B. Tleulessova
e-mail: agila72@mail.ru

A. B. Tleulessova
L.N. Gumilyov Eurasian National University,
13 Kazhymukhan str., Z01C0X0 Nur-Sultan, Kazakhstan

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_7

and some types of boundary value problems for impulsive differential equations were studied (see and their references) in [1–6, 13, 18, 27]. Periodic and nonlocal problems for the impulsive system of partial differential equations arise from mathematical modeling of numerous processes in biology, physics, chemistry, mechanics. Periodic and some other types of nonlocal boundary value problems for impulsive hyperbolic equations were studied in [12, 14, 19, 23, 26, 28]. To study the solvability of these classes of problems there have been applied the methods of the qualitative theory of differential equations and oscillations theory, Riemann's method, the numerical-analytical method, the monotone iteration method, asymptotic methods, the upper and lower solutions method, and others. Nevertheless, the problem of finding effective features for the unique solvability of nonlocal problems for the impulsive system of higher order partial differential equations is still actual today.

The aim of this paper is to establish the conditions for the existence and uniqueness of the solution to the impulsive system of fourth order partial differential equations.

In this Section, we give the statement of the problem and the assumptions regarding the initial data.

On the domain $\Omega = [0, T] \times [0, \omega]$, consider the following problem:

$$\partial_x^3 \partial_t u = A_1(t,x)\partial_x^3 u + B_1(t,x)\partial_x^2 \partial_t u + A_2(t,x)\partial_x^2 u$$

$$+ B_2(t,x)\partial_x^2 \partial_t u + A_3(t,x)\partial_x u + B_3(t,x)\partial_t u + C(t,x)u + f(t,x), \tag{1}$$

$$P(x)\partial_x^2 \partial_t u(t,x)\big|_{t=0} + S(x)\partial_x^2 \partial_t u(t,x)\big|_{t=T} = \varphi_0(x), \qquad x \in [0,\omega], \tag{2}$$

$$\lim_{t \to t_r+0} \partial_x^2 \partial_t u(t,x) - \lim_{t \to t_r-0} \partial_x^2 \partial_t u(t,x) = \varphi_r(x), \quad r = 1, 2, ..., k, \quad x \in [0,\omega], \tag{3}$$

$$u(t,0) = \psi_0(t), \quad \partial_x u(t,x)\big|_{x=0} = \psi_1(t), \quad \partial_x^2 u(t,x)\big|_{x=0} = \psi_2(t), \ t \in [0,T], \tag{4}$$

where $u(t,x) = col(u_1(t,x), u_2(t,x), ..., u_n(t,x))$ is an unknown function, $\partial_t u(t,x) = \frac{\partial u(t,x)}{\partial t}$, $\partial_x^i u(t,x) = \frac{\partial^i u(t,x)}{\partial x^i}$, $\partial_x^i \partial_t u(t,x) = \frac{\partial^{i+1} u(t,x)}{\partial x^i \partial t}$, $i = 1, 2, 3$,

$0 < t_1 < t_2 < ... < t_k < T$; the $n \times n$ matrices $A_i(t,x)$, $B_i(t,x)$, $i = 1, 2, 3$, $C(t,x)$, and n vector function $f(t,x)$ are piecewise continuous on Ω with possible discontinuities at the lines $t = t_r, r = 1, 2, ..., k$; the $n \times n$ matrices $P(x)$, $S(x)$, and n vector function $\varphi_0(x)$ are continuously differentiable on $[0,\omega]$; the n vector-functions $\psi_i(t)$, $i = 0, 1, 2$, are continuous on $[0, T]$ and piecewise continuously differentiable on $[0, T]$ with possible discontinuities at the lines $t = t_r$, $r = 1, 2, ..., k$; the n vector functions $\varphi_r(x)$, $r = 1, 2, ..., k$, are continuously differentiable on $[0,\omega]$.

The following compatibility conditions regarding initial data hold:

$$P(0)\dot{\psi}_2(0) + S(0)\dot{\psi}_2(T) = \varphi_0(0), \tag{5}$$

$$\lim_{t \to t_r+0} \dot{\psi}_2(t) - \lim_{t \to t_r-0} \dot{\psi}_2(t) = \varphi_r(0), \quad r = 1, 2, ..., k. \tag{6}$$

Introduce notation
$t_0 = 0, t_{k+1} = T, \Omega_r = [t_{r-1}, t_r) \times [0, \omega], r = 1, 2, ..., k+1,$ i.e. $\Omega = \bigcup_{r=1}^{k+1} \Omega_r$.
Let $PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$ be the space of vector functions $u : \Omega \to \mathscr{R}^n$ continuous on Ω with possible discontinuities at lines $t = t_r$, with the norm

$$||u||_1 = \max_{r=1,2,...,k+1} \sup_{(t,x)\in\Omega_r} ||u(t,x)||.$$

Note, for all $x \in [0, \omega]$, there exist the left-handed limits $\lim\limits_{t\to t_r-0} u(t,x)$ and continuous right-handed limits $\lim\limits_{t\to t_r+0} u(t,x)$ at $t = t_r$, $r = 1, 2, ..., k$.

A function $u(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$ with partial derivatives

$$\partial_x^i \partial_t^j u(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n), \qquad i = 1, 2, 3, \qquad j = 0, 1,$$

is said to be a *solution* to problem (1)–(4) if it satisfies system (1) for all $(t,x) \in \Omega$, except the lines $t = t_r, r = 1, 2, ..., k$, and the boundary conditions (2), the conditions of impulse effects at fixed times (3) and conditions (4).

Various problems for different classes of fourth order partial differential equations are studied in [15–17, 20–22, 24, 25].

We investigate the existence and uniqueness of solution to problem (1)–(4) by the method of introducing additional parameters [7–11].

2 Scheme of the Method

In this Section, by method of introducing additional parameters [7–11] we reduce the original problem (1)–(4) to an equivalent nonlocal problem for an impulsive system of second order hyperbolic equations and integral relations.

Introduce new unknown functions
$v_1(t,x) = \partial_x^2 u(t,x), \quad v_2(t,x) = \partial_x u(t,x), \quad v_3(t,x) = u(t,x).$
Taking into account first and second conditions of (4), we have:

$$v_2(t,x) = \psi_1(t) + \int_0^x v_1(t,\xi)d\xi,$$

$$v_3(t,x) = \psi_0(t) + \psi_1(t)x + \int_0^x (x-\xi)v_1(t,\xi)d\xi.$$

Then reduce problem (1)–(4) to the following problem:

$$\partial_x\partial_t v_1 = A_1(t,x)\partial_x v_1 + B_1(t,x)\partial_t v_1 + A_2(t,x)v_1 + f(t,x) + F(t,x,v_2,v_3), \tag{7}$$

$$P(x)\partial_t v_1(t,x)\big|_{t=0}+S(x)\partial_t v_1(t,x)\big|_{t=T}=\varphi_0(x), \tag{8}$$

$$\lim_{t\to t_r+0}\partial_t v_1(t,x)-\lim_{t\to t_r-0}\partial_t v_1(t,x)=\varphi_r(x), \qquad x\in[0,\omega], \qquad r=1,2,...,k, \tag{9}$$

$$v_1(t,0)=\psi_2(t), \qquad t\in[0,T], \tag{10}$$

$$v_2(t,x)=\psi_1(t)+\int_0^x v_1(t,\xi)d\xi, \tag{11}$$

$$v_3(t,x)=\psi_0(t)+\psi_1(t)x+\int_0^x (x-\xi)v_1(t,\xi)d\xi, \tag{12}$$

where
$F(t,x,v_2,v_3)=A_3(t,x)v_2(t,x)+B_2(t,x)\partial_t v_2(t,x)+B_3(t,x)\partial_t v_3(t,x)+C(t,x)v_3(t,x)$.

Differentiating relations (11), (12) by t, we obtain the following equalities for partial derivatives $\partial_t v_s(t,x)$:

$$\partial_t v_2(t,x)=\dot{\psi}_1(t)+\int_0^x \partial_t v_1(t,\xi)d\xi, \tag{13}$$

$$\partial_t v_3(t,x)=\dot{\psi}_0(t)+\dot{\psi}_1(t)x+\int_0^x (x-\xi)\partial_t v_1(t,\xi)d\xi. \tag{14}$$

A system of 3 vector functions $(v_1(t,x), v_2(t,x), v_3(t,x))$, where
$v_1(t,x)\in PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$, $\partial_x v_1(t,x)$, $\partial_t v_1(t,x)$, $\partial_x\partial_t v_1(t,x)\in PC(\Omega,$ $\{t_r\}_{r=1}^k,\mathscr{R}^n)$, and $v_s(t,x), \partial_t v_s(t,x)\in PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$, $s=2,3$,
is said to be a solution to problem (7)–(12), if it satisfies the impulsive system of second order hyperbolic equations (7) for all $(t,x)\in\Omega$, except the lines $t=t_r$, $r=1,2,...,k$, the boundary conditions (8), (10), conditions of impulse effects at fixed times (9), and integral relations (11), (12). Here the functions $v_2(t,x)$ and $v_3(t,x)$ are connected with function $v_1(t,x)$ by integral conditions (11) and (12), respectively.

Problem (1)–(4) is equivalent to problem (7)–(12).

Let $u^*(t,x)$ be a solution to nonlocal problem (1)–(4). Then the system of 3 vector functions $(v_1^*(t,x), v_2^*(t,x), v_3^*(t,x))$, where
$v_1^*(t,x)=\partial_x^2 u^*(t,x)$, $v_2^*(t,x)=\partial_x u^*(t,x)$, $v_3^*(t,x)=u^*(t,x)$,
is a solution to problem (7)–(12). Conversely, if a system of 3 vector functions $(\widetilde{v}_1(t,x), \widetilde{v}_2(t,x), \widetilde{v}_3(t,x))$ is the solution to problem (7)–(12), then $\widetilde{u}(t,x)$ defined by equality

$$\widetilde{u}(t,x)=\psi_0(t)+\psi_1(t)x+\int_0^x (x-\xi)\widetilde{v}_1(t,\xi)d\xi$$

is a solution to nonlocal problem (1)–(4).

At fixed $v_2(t,x)$ and $v_3(t,x)$, problem (7)–(12) is a nonlocal problem for impulsive system of second order hyperbolic equations with respect to $v_1(t,x)$ on Ω. The integral relations (11) and (12) allow us to determine the unknown functions $v_2(t,x)$ and $v_3(t,x)$, respectively. From (13) and (14) we define the partial derivatives $\partial_t v_s(t,x)$, $s=2,3$.

The problem (7)–(12) can be interpreted as:

- a nonlocal problem for the impulsive system of second order hyperbolic equations with distributed parameters $v_s(t,x)$, $s=2,3$.
- an inverse problem for impulsive system of second order hyperbolic equations, where the unknown functions $v_s(t,x)$, $s=2,3$, are determined from integral relations (11), (12).
- a control problem for the impulsive system of second order hyperbolic equations, where the control functions $v_s(t,x)$, $s=2,3$, satisfy integral constraints (11), (12).

Hereby, problem (1)–(4) is reduced to an equivalent nonlocal problem for impulsive system of second order hyperbolic equations with parameters and integral conditions.

3 Algorithm for Finding a Solution to Problem (7)–(12)

In this Section, we propose an algorithm for finding the approximate solution to problem (7)–(12).

If we know $v_1(t,x)$ and its derivatives $\partial_x v_1(t,x)$, $\partial_t v_1(t,x)$, then from (11)–(14) we find $v_2(t,x)$, $v_3(t,x)$ and their partial derivatives $\partial_t v_2(t,x)$, $\partial_t v_3(t,x)$, respectively. Conversely, if we know $v_2(t,x)$, $\partial_t v_2(t,x)$, $v_3(t,x)$, $\partial_t v_3(t,x)$, then from (7)–(10) we can find $v_1(t,x)$ and its partial derivatives $\partial_x v_1(t,x)$, $\partial_t v_1(t,x)$.

Since the function $v_1(t,x)$ and the functions $v_2(t,x)$, $v_3(t,x)$ are unknown, to find a solution to problem (7)–(12) we use an iterative method.

We determine a system of 3 vector functions $(v_1^*(t,x), v_2^*(t,x), v_3^*(t,x))$ as a limit of system of 3 vector functions $(v_1^{(m)}(t,x), v_2^{(m)}(t,x), v_3^{(m)}(t,x))$, $m=0,1,2,...$, by the following algorithm:

Step - 0. 1) In right-hand side of system (7), set $v_2(t,x)=\psi_1(t)$, $v_3(t,x)=\psi_0(t)+\psi_1(t)x$, $\partial_t v_2(t,x)=\dot{\psi}_1(t)$, $\partial_t v_3(t,x)=\dot{\psi}_0(t)+\dot{\psi}_1(t)x$.

Then from nonlocal problem for the impulsive system of hyperbolic equations (7)–(10) we find $v_1^{(0)}(t,x)$ for all $(t,x)\in\Omega$. We also find its partial derivatives $\partial_x v_1^{(0)}(t,x)$, $\partial_t v_1^{(0)}(t,x)$ and $\partial_x\partial_t v_1^{(0)}(t,x)$ for all $(t,x)\in\Omega$.

2) From integral relations (11)–(14) we determine $v_s^{(0)}(t,x)$ and $\partial_t v_s^{(0)}(t,x)$, $s=2,3$, for all $(t,x)\in\Omega$:

$$v_2^{(0)}(t,x)=\psi_1(t)+\int_0^x v_1^{(0)}(t,\xi)d\xi,$$

$$v_3^{(0)}(t,x) = \psi_0(t) + \psi_1(t)x + \int_0^x (x-\xi)v_1^{(0)}(t,\xi)d\xi,$$
$$\partial_t v_2^{(0)}(t,x) = \dot{\psi}_1(t) + \int_0^x \partial_t v_1^{(0)}(t,\xi)d\xi,$$
$$\partial_t v_3^{(0)}(t,x) = \dot{\psi}_0(t) + \dot{\psi}_1(t)x + \int_0^x (x-\xi)\partial_t v_1^{(0)}(t,\xi)d\xi.$$

Step - 1. 1) In right-hand side of system (7), suppose that $v_2(t,x) = v_2^{(0)}(t,x)$, $v_3(t,x) = v_3^{(0)}(t,x)$, $\partial_t v_2(t,x) = \partial_t v_2^{(0)}(t,x)$, $\partial_t v_3(t,x) = \partial_t v_3^{(0)}(t,x)$, .

Then from nonlocal problem for impulsive system of hyperbolic equations (7)–(10) we find $v_1^{(1)}(t,x)$ for all $(t,x) \in \Omega$. We find its partial derivatives $\partial_x v_1^{(1)}(t,x)$, $\partial_t v_1^{(1)}(t,x)$ and $\partial_x \partial_t v_1^{(1)}(t,x)$ as well for all $(t,x) \in \Omega$.

2) From integral relations (11)–(14) we determine $v_s^{(1)}(t,x)$ and $\partial_t v_s^{(1)}(t,x)$, $s = 2, 3$, for all $(t,x) \in \Omega$:

$$v_2^{(1)}(t,x) = \psi_1(t) + \int_0^x v_1^{(1)}(t,\xi)d\xi,$$
$$v_3^{(1)}(t,x) = \psi_0(t) + \psi_1(t)x + \int_0^x (x-\xi)v_1^{(1)}(t,\xi)d\xi,$$
$$\partial_t v_2^{(1)}(t,x) = \dot{\psi}_1(t) + \int_0^x \partial_t v_1^{(1)}(t,\xi)d\xi,$$
$$\partial_t v_3^{(1)}(t,x) = \dot{\psi}_0(t) + \dot{\psi}_1(t)x + \int_0^x (x-\xi)\partial_t v_1^{(1)}(t,\xi)d\xi.$$

And so on.

Step - m. 1) In right-hand side of system (7), suppose that $v_2(t,x) = v_2^{(m-1)}(t,x)$, $v_3(t,x) = v_2^{(m-1)}(t,x)$, $\partial_t v_2(t,x) = \partial_t v_2^{(m-1)}(t,x)$, $\partial_t v_3(t,x) = \partial_t v_3^{(m-1)}(t,x)$.

Then from nonlocal problem for impulsive system of hyperbolic equations (7)–(10) we find $v_1^{(m)}(t,x)$ for all $(t,x) \in \Omega$. We find its partial derivatives $\partial_x v_1^{(m)}(t,x)$, $\partial_t v_1^{(m)}(t,x)$ and $\partial_x \partial_t v_1^{(m)}(t,x)$ as well for all $(t,x) \in \Omega$.

2) From integral relations (11)–(14) we determine $v_s^{(m)}(t,x)$ and $\partial_t v_s^{(m)}(t,x)$, $s = 2, 3$, for all $(t,x) \in \Omega$:

$$v_2^{(m)}(t,x) = \psi_1(t) + \int_0^x v_1^{(m)}(t,\xi)d\xi,$$
$$v_3^{(m)}(t,x) = \psi_0(t) + \psi_1(t)x + \int_0^x (x-\xi)v_1^{(m)}(t,\xi)d\xi,$$
$$\partial_t v_2^{(m)}(t,x) = \dot{\psi}_1(t) + \int_0^x \partial_t v_1^{(m)}(t,\xi)d\xi,$$
$$\partial_t v_3^{(m)}(t,x) = \dot{\psi}_0(t) + \dot{\psi}_1(t)x + \int_0^x (x-\xi)\partial_t v_1^{(m)}(t,\xi)d\xi.$$

$m = 2, 3, ...$

4 Unique Solvability of the Nonlocal Problem for Impulsive System of Fourth Order Partial Differential Equations

Consider an auxiliary nonlocal problem for impulsive system of second order hyperbolic equations

$$\partial_x \partial_t v_1 = A_1(t,x)\partial_x v_1 + B_1(t,x)\partial_t v_1 + A_2(t,x)v_1 + F(t,x), \tag{15}$$

$$P(x)\partial_t v_1(t,x)\big|_{t=0} + S(x)\partial_t v_1(t,x)\big|_{t=T} = \varphi_0(x), \tag{16}$$

$$\lim_{t\to t_r+0} \partial_t v_1(t,x) - \lim_{t\to t_r-0} \partial_t v_1(t,x) = \varphi_r(x), \qquad x\in[0,\omega], \qquad r=1,2,...,k; \tag{17}$$

$$v_1(t,0) = \psi_2(t), \qquad t\in[0,T]. \tag{18}$$

Here the function $F(t,x)$ belongs to $PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$.

A vector function $v_1(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$, having partial derivatives $\partial_x v_1(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$, $\partial_t v_1(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$, $\partial_x\partial_t v_1(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$ is said to be a *solution* to problem (15)–(18) if it satisfies system (15) for all $(t,x)\in\Omega$, except the lines $t=t_r$, $r=1,2,...,k$ and meets the conditions (16), (18) and conditions of impulse effects at fixed times (17).

Conditions (16) and (17) include the values of partial derivatives of desired function by time variable t.

The impulsive system (15) with boundary conditions (16), (18) and conditions of impulse effects at fixed times (17) is considered for the first time.

The impulsive system (15) with various type conditions (including the values of desired function or the values of partial derivatives of desired function by spatial variable x) are studied in [7–11]. Sufficient conditions for the unique solvability of the considered problem are established in the terms of hyperbolic system coefficients, boundary and impulsive matrices. For example, in [10], we establish the conditions for unique solvability of system (15) with condition (18) and the following conditions:

$$P(x)\partial_x v_1(t,x)\big|_{t=0} + S(x)\partial_x v_1(t,x)\big|_{t=T} = \varphi_0(x),$$

$$\lim_{t\to t_r+0} \partial_x v_1(t,x) - \lim_{t\to t_r-0} \partial_x v_1(t,x) = \varphi_r(x), \qquad r=1,2,...,k.$$

Next assertion provides us the feasibility and convergence of the proposed algorithm.

Theorem 1. *Suppose*

(i) the $n \times n$ matrices $A_i(t,x)$, $B_i(t,x)$, $i = 1, 2, 3$, $C(t,x)$, and n vector function $f(t,x)$ are piecewise continuous on Ω with possible discontinuities at lines $t = t_r$, $r = 1, 2, ..., k$;

(ii) the $n \times n$ matrices $P(x)$, $S(x)$, n vector functions $\varphi_r(x)$, $r = 0, 1, 2, ..., k$ are continuously differentiable on $[0, \omega]$;

(iii) the n vector-functions $\psi_i(t)$, $i = 0, 1, 2$ are continuous on $[0, T]$ and piecewise continuously differentiable on $[0, T]$ with possible discontinuities at lines $t = t_r$, $r = 1, 2, ..., k$ and satisfy compatibility conditions (5)–(6).

(iv) nonlocal problem (15)–(18) is uniquely solvable for any $F(t,x) \in PC(\Omega, \{t_r\}_{r=1}^k, \mathscr{R}^n)$, $\varphi_i(x) \in C^1([0,\omega], \mathscr{R}^n)$, $i = 0, 1, 2, ..., k$, and $\psi_2(t) \in PC^1([0,T], \{t_r\}_{r=1}^k, \mathscr{R}^n)$.

Then problem (7)–(12) has a unique solution.

Proof. Let conditions (i)–(iv) of Theorem be valid. Then nonlocal problem (15)–(18) has a unique solution. Applying the algorithm, we will find a solution to problem (7)–(12). From step zero step of the algorithm we find a solution to the problem

$$\partial_x \partial_t v_1 = A_1(t,x)\partial_x v_1 + B_1(t,x)\partial_t v_1 + A_2(t,x)v_1 + f(t,x) + F^{(0)}(t,x,v_2,v_3), \tag{19}$$

$$P(x)\partial_t v_1(t,x)\Big|_{t=0} + S(x)\partial_t v_1(t,x)\Big|_{t=T} = \varphi_0(x), \tag{20}$$

$$\lim_{t \to t_r+0} \partial_t v_1(t,x) - \lim_{t \to t_r-0} \partial_t v_1(t,x) = \varphi_r(x), \qquad x \in [0,\omega], \qquad r = 1, 2, ..., k; \tag{21}$$

$$v_1(t,0) = \psi_2(t), \qquad t \in [0,T], \tag{22}$$

where

$$F^{(0)}(t,x,v_2,v_3) = A_3(t,x)\psi_1(t) + B_2(t,x)\dot{\psi}_1(t) + B_3(t,x)[\dot{\psi}_0(t) + \dot{\psi}_1(t)x] + C(t,x)[\psi_0(t) + \psi_1(t)x].$$

Assume

$$||v_1(\cdot,x)||_1 = \max_{r=1,2,...,k+1} \sup_{t \in [t_{r-1},t_r)} ||v_1(t,x)||,$$

$$\Phi_r(x) = \max\Big(||\varphi_r(x)||, ||\dot{\varphi}_r(x)||\Big), \qquad r = 1, 2, ..., k;$$

$$\Psi_l = \max\Big(\max_{t \in [0,T]} ||\psi_l(t)||, \max_{r=1,2,...,k+1} \sup_{t \in [t_{r-1},t_r)} ||\dot{\psi}_l(t)||\Big), \qquad l = 0, 1, 2.$$

By assumption, problem (19)–(22) has a unique solution $v_1^{(0)}(t,x)$ satisfying the following estimate:

$$\max\Big(||v_1^{(0)}(\cdot,x)||_1,\, ||\partial_x v_1^{(0)}(\cdot,x)||_1,\, ||\partial_t v_1^{(0)}(\cdot,x)||_1\Big)$$

$$\leq K(1+K_1)\max\Big(||f(\cdot,x)||_1,\, ||\varphi_0(x)||,\, \max_{r=1,2,\ldots,k} \Phi_r(x),\, \max_{l=0,1,2} \Psi_l\Big), \tag{23}$$

where K is a constant defined via $\alpha_1(x) = \max_{r=\overline{1,k+1}} \sup_{t\in[t_{r-1},t_r)} ||A_1(t,x)||$,
$h = \max_{r=1,2,\ldots,k+1}(t_r - t_{r-1})$, and $P(x)$, $S(x)$ [8],
$K_1 = \max\Big[\sum_{s=2}^{3}\big\{||A_s||_1 + ||B_s||_1\big\} + ||C||_1,\, \max_{x\in[0,\omega]}\big\{||P(x)|| + ||S(x)||\big\}\Big]$.
Then from integral relations (11)–(14), we have:

$$v_2^{(0)}(t,x) = \psi_1(t) + \int_0^x v_1^{(0)}(t,\xi)d\xi, \tag{24}$$

$$v_3^{(0)}(t,x) = \psi_0(t) + \psi_1(t)x + \int_0^x (x-\xi)v_1^{(0)}(t,\xi)d\xi. \tag{25}$$

$$\partial_t v_2^{(0)}(t,x) = \dot{\psi}_1(t) + \int_0^x \partial_t v_1^{(0)}(t,\xi)d\xi, \tag{26}$$

$$\partial_t v_3^{(0)}(t,x) = \dot{\psi}_0(t) + \dot{\psi}_1(t)x + \int_0^x (x-\xi)\partial_t v_1^{(0)}(t,\xi)d\xi. \tag{27}$$

Let $v_s^{(m-1)}(t,x)$, $s=2,3$, be given.

Then solving problem (7)–(10) for $v_s(t,x) = v_s^{(m-1)}(t,x), s=2,3$, we find function $v_1^{(m)}(t,x)$, $m=1,2,\ldots$.

For founded $v_1^{(m}(t,x)$, we determine next approximations $v_s(t,x)$, $s=2,3$, from relations (11)–(14):

$$v_2^{(m)}(t,x) = \psi_1(t) + \int_0^x v_1^{(m)}(t,\xi)d\xi, \tag{28}$$

$$v_3^{(m)}(t,x) = \psi_0(t) + \psi_1(t)x + \int_0^x (x-\xi)v_1^{(m)}(t,\xi)d\xi. \tag{29}$$

$$\partial_t v_2^{(m)}(t,x) = \dot{\psi}_1(t) + \int_0^x \partial_t v_1^{(m)}(t,\xi)d\xi, \tag{30}$$

$$\partial_t v_3^{(m)}(t,x) = \dot{\psi}_0(t) + \dot{\psi}_1(t)x + \int_0^x (x-\xi)\partial_t v_1^{(m)}(t,\xi)d\xi. \tag{31}$$

Compose the differences $\Delta v_1^{(m)}(t,x) = v_1^{(m)}(t,x) - v_1^{(m-1)}(t,x)$, $\Delta v_2^{(m)}(t,x) = v_2^{(m)}(t,x) - v_2^{(m-1)}(t,x)$, and $\Delta v_3^{(m)}(t,x) = v_3^{(m)}(t,x) - v_3^{(m-1)}(t,x)$.

Using the unique solvability of problem (15)–(18), we establish the following estimates:

$$\max\Big(||\Delta v_1^{(m+1)}(\cdot,x)||_1, ||\partial_x \Delta v_1^{(m+1)}(\cdot,x)||_1, ||\partial_t \Delta v_1^{(m+1)}(\cdot,x)||_1\Big)$$

$$\leq K\cdot(1+K_1)\max\Big(\max_{s=2,3}||\Delta v_s^{(m)}(\cdot,x)||_1, \max_{s=2,3}||\partial_t \Delta v_s^{(m)}(\cdot,x)||_1\Big), \tag{32}$$

$$\max\Big(||\Delta v_2^{(m)}(\cdot,x)||_1, ||\partial_t \Delta v_2^{(m)}(\cdot,x)||_1\Big)$$

$$\leq \int_0^x \max\Big(||\Delta v_1^{(m)}(\cdot,\xi)||_1, ||\partial_x \Delta v_1^{(m)}(\cdot,\xi)||_1, ||\partial_t \Delta v_1^{(m)}(\cdot,\xi)||_1\Big)d\xi, \tag{33}$$

$$\max\Big(||\Delta v_3^{(m)}(\cdot,x)||_1, ||\partial_t \Delta v_3^{(m)}(\cdot,x)||_1\Big)$$

$$\leq \int_0^x (x-\xi)\max\Big(||\Delta v_1^{(m)}(\cdot,\xi)||_1, ||\partial_x \Delta v_1^{(m)}(\cdot,\xi)||_1, ||\partial_t \Delta v_1^{(m)}(\cdot,\xi)||_1\Big)d\xi, \tag{34}$$

$m = 1, 2, \ldots.$

Inequality (34) yields the main inequality:

$$\max\Big(||\Delta v_1^{(m+1)}(\cdot,x)||_1, ||\partial_x \Delta v_1^{(m+1)}(\cdot,x)||_1, ||\partial_t \Delta v_1^{(m+1)}(\cdot,x)||_1\Big)$$

$$\leq K\cdot(1+K_1)\int_0^x \max[1,(x-\xi)]$$

$$\times \max\Big(||\Delta v_1^{(m)}(\cdot,\xi)||_1, ||\partial_x \Delta v_1^{(m)}(\cdot,\xi)||_1, ||\partial_t \Delta v_1^{(m)}(\cdot,\xi)||_1\Big)d\xi, \tag{35}$$

where $x \in [0,\omega]$.

Inequality (35) provides the uniform convergence of sequences $\{v_1^{(m)}(t,x)\}$ and $\left\{\partial_t v_1^{(m)}(t,x)\right\}$ in $PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$ as $m\to\infty$. Then the uniform convergence of sequences $\{v_s^{(m)}(t,x)\}$ and $\left\{\partial_t v_s^{(m)}(t,x)\right\}$, $s=2,3$ on Ω as $m\to\infty$ follows from (33) and (34). The limit functions $v_1^*(t,x)$, $\partial_t v_1^*(t,x)$, $v_s^*(t,x)$, and $\partial_t v_s^*(t,x)$, $s=2,3$, belong to $PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$. The system of 3 functions $\{v_1^*(t,x), v_2^*(t,x), v_3^*(t,x)\}$ is a solution to problem (7)–(12). Theorem 1 is proved. □

Therefore, Theorem 1 provides the unique solvability of problem (7)–(12) in the terms of initial data at unique solvability of auxiliary nonlocal problem (15)–(18).

The equivalence of problems (7)–(12) and (1)–(4) yields the following assertion.

Theorem 2. *Under conditions (i)–(iv) of Theorem 1, the original nonlocal problem for the impulsive system of fourth order partial differential equations (1)–(4) has a unique solution.*

So, the unique solvability of the auxiliary nonlocal problem for the impulsive system of second order hyperbolic equations (15)–(18) is the main condition for unique solvability of problem (1)–(4).

5 The Periodic Problem for an Impulsive System of Fourth Order Partial Differential Equations

In this Section, as an example, we consider the periodic problem for the impulsive system of the fourth order partial differential equations. We formulate the results of Sect. 4 for a periodical case of problem (1)–(4):
$P(x)=I$, $S(x)=-I$, where I is the identity matrix on dimension n, and $\varphi_0(x)=0$.

On the domain $\Omega=[0,T]\times[0,\omega]$, we consider a periodic problem for the impulsive system of fourth order partial differential equations

$$\partial_x^3\partial_t u = A_1(t,x)\partial_x^3 u + B_1(t,x)\partial_x^2\partial_t u + A_2(t,x)\partial_x^2 u$$

$$+ B_2(t,x)\partial_x\partial_t u + A_3(t,x)\partial_x u + B_3(t,x)\partial_t u + C(t,x)u + f(t,x), \tag{36}$$

$$\partial_x^2\partial_t u\Big|_{t=0} = \partial_x^2\partial_t u\Big|_{t=T}, \qquad x\in[0,\omega], \tag{37}$$

$$\lim_{t\to t_r+0}\partial_x^2\partial_t u - \lim_{t\to t_r-0}\partial_x^2\partial_t u = \varphi_r(x), \quad r=1,2,...,k; \tag{38}$$

$$u(t,0)=\psi_0(t),\ \partial_x u(t,x)\Big|_{x=0}=\psi_1(t),\ \partial_x^2 u(t,x)\Big|_{x=0}=\psi_2(t), \qquad t\in[0,T]. \tag{39}$$

A function $u(t,x)\in PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$ with partial derivatives

$$\partial_x^i\partial_t^j u(t,x)\in PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n), \qquad i=1,2,3, \quad j=0,1,$$

is said to be a *solution* to problem (36)–(39) if it satisfies system (36) for all $(t,x)\in\Omega$, except the lines $t=t_r$, $r=1,2,...,k$, and the periodic condition (37), the conditions of impulse effects at fixed times (38) and conditions (39).

Introduce new unknown functions:

$v_1(t,x)=\partial_x^2 u(t,x),\quad v_2(t,x)=\partial_x u(t,x),\quad v_3(t,x)=u(t,x).$

Then reduce periodic problem (36)–(39) to an equivalent problem:

$$\partial_x\partial_t v_1 = A_1(t,x)\partial_x v_1 + B_1(t,x)\partial_t v_1 + A_2(t,x)v_1 + f(t,x) + F(t,x,v_2,v_3), \tag{40}$$

$$\partial_t v_1(t,x)\big|_{t=0} = \partial_t v_1(t,x)\big|_{t=T}, \qquad x\in[0,\omega], \tag{41}$$

$$\lim_{t\to t_r+0}\partial_t v_1(t,x) - \lim_{t\to t_r-0}\partial_t v_1(t,x) = \varphi_r(x), \qquad r=1,2,...,k; \tag{42}$$

$$v_1(t,0)=\psi_2(t), \qquad t\in[0,T], \tag{43}$$

$$v_2(t,x)=\psi_1(t)+\int_0^x v_1(t,\xi)d\xi, \tag{44}$$

$$v_3(t,x)=\psi_0(t)+\psi_1(t)x+\int_0^x (x-\xi)v_1(t,\xi)d\xi, \tag{45}$$

where $F(t,x,v_2,v_3)=A_3(t,x)v_2(t,x)+\sum\limits_{s=2}^{3}B_s(t,x)\partial_t v_s(t,x)+C(t,x)v_3(t,x).$

A system of 3 vector functions $(v_1(t,x),v_2(t,x),v_3(t,x))$ with $v_s(t,x)$, $\frac{\partial v_s(t,x)}{\partial t}\in PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$, $s=\overline{1,3}$, and $\frac{\partial v_1(t,x)}{\partial x}$, $\frac{\partial^2 v_1(t,x)}{\partial x\partial t}\in PC(\Omega,\{t_r\}_{r=1}^k,\mathscr{R}^n)$, is said to be a *solution* to periodic problem (40)–(45), if it satisfies the impulsive system of second order hyperbolic equations (40) for all $(t,x)\in\Omega$, except the lines $t=t_r$, $r=1,2,...,k$, the periodic condition (41), and conditions of impulse effects at fixed times (42), condition (43) and integral relations (44), (45). Functions $v_2(t,x)$ and $v_3(t,x)$ are connected with function $v_1(t,x)$ by integral conditions (44) and (45), respectively.

Theorem 3. *Suppose*

(i) the $n\times n$ matrices $A_j(t,x)$, $B_j(t,x)$, $j=1,2,3$, $C(t,x)$, and n vector function $f(t,x)$ are piecewise continuous on Ω with possible discontinuities at lines $t=t_r$, $r=1,2,...,k$;

(ii) the n vector functions $\varphi_r(x)$, $r=1,2,...,k$ are continuously differentiable on $[0,\omega]$*; the n vector-functions $\psi_s(t)$, $s=0,1,2$ are continuous on* $[0,T]$ *and piecewise continuously differentiable on* $[0,T]$ *with possible discontinuities at lines $t=t_r$, $r=1,2,...,k$, and satisfy the following compatibility conditions:*

$$\dot\psi_2(0)=\dot\psi_2(T), \qquad \lim_{t\to t_r+0}\dot\psi_2(t)-\lim_{t\to t_r-0}\dot\psi_2(t)=\varphi_r(0), \qquad r=1,2,...,k;$$

(iii) the periodic problem (15), (17), (18) with condition (41) is uniquely solvable for any $F(t, x) \in PC(\Omega, \{t_r\}_{r=1}^{k}, \mathscr{R}^n)$, $\varphi_r(x) \in C([0, \omega], \mathscr{R}^n)$, $r = 1, 2, ..., k$, and $\psi_2(t) \in PC^1([0, T], \mathscr{R}^n)$.

Then periodic problem (36)–(39) has a unique solution.

The proof of Theorem 3 is similar to the proof of Theorem 1.

Acknowledgements One of the article authors (Assanova) has participated at the International Conference on Differential & Difference Equations and Applications 2019 in Lisbon.

We sincerely thank Prof. Sandra Pinelas, the Chair of the conference, for the excellent organization of ICDDEA-2019.

Researches presented in the article are supported by the grant of the Ministry of Education and Science of the Republic of Kazakhstan, Project No. AP 05131220, for 2018–2020 years.

The authors thank the referees for his/her careful reading of the manuscript and useful suggestions which allowed us to improve the present paper.

References

1. Akhmet, M.U.: Principles of Discontinuous Dynamical Systems. Springer, New York (2010)
2. Akhmet, M.U.: Nonlinear Hybrid Continuous/Discrete-Time Models. Atlantis Press, Paris (2011)
3. Akhmet, M.: Almost Periodicity, Chaos, and Asymptotic Equivalence. Springer, New York (2020)
4. Akhmet, M., Fen, M.O.: Replication of Chaos in Neural Networks, Economics and Physics. Nonlinear Physical Science. Springer, Higher Education Press, Beijing, Heidelberg (2016)
5. Akhmet, M., Kashkynbayev, A.: Bifurcation in Autonomous and Nonautonomous Differential Equations with Discontinuities. Springer, Higher Education Press, Heidelberg (2017)
6. Akhmet, M.U., Yilmaz, E.: Neural Networks with Discontinuous/Impact Activations. Springer, New York (2013)
7. Asanova, A.T.: On a nonlocal boundary-value problem for systems of impulsive hyperbolic equations. Ukr. Math. J. **65**(3), 349–365 (2013)
8. Asanova, A.T.: Well-posed solvability of a nonlocal boundary-value problem for systems of hyperbolic equations with impulse effects. Ukr. Math. J. **67**(3), 333–346 (2015)
9. Assanova, A.T.: On the solvability of nonlocal boundary value problem for the systems of impulsive hyperbolic equations with mixed derivatives. J. Discontinuity Nonlinearity Complexity **5**(2), 153–165 (2016)
10. Asanova, A.T., Kadirbaeva, Z.M., Bakirova, E.A.: About of an unique solvability of a nonlocal boundary value problem for the loaded systems of hyperbolic equations with impulse effects. Ukr. Math. J. **69**(8), 1175–1195 (2018)
11. Assanova, A.T., Kadirbayeva, Z.M.: Periodic problem for an impulsive system of the loaded hyperbolic equations. Electron. J. Differ. Equ. **2018**(72), 1–8 (2018)
12. Bainov, D.D., Minchev, E., Myshkis, A.: Periodic boundary value problems for impulsive hyperbolic systems. Commun. Appl. Anal. **1**(4), 1–14 (1997)
13. Bainov, D.D., Simeonov, P.S.: Systems with Impulse Effect: Stability, Theory and Applications. Halsted Press, New York - Chichester - Brisbane - Toronto (1989)
14. Belarbi, A., Benchohra, M.: Existence theory for perturbed impulsive hyperbolic differential inclusions with variable times. J. Math. Anal. Appl. **327**, 1116–1129 (2007)
15. Ferraioli, D.C., Tenenblat, K.: Fourth order evolution equations which describe pseudospherical surfaces. J. Differ. Equ. **257**, 3165–3199 (2014)

16. Kiguradze, T.: On solvability and well-posedness of boundary value problems for nonlinear hyperbolic equations of the fourth order. Georgian Math. J. **15**(3), 555–569 (2008)
17. Kiguradze, T., Lakshmikantham, V.: On the Dirichlet problem for fourth order linear hyperbolic equations. Nonlinear Anal. **49**(2), 197–219 (2002)
18. Lakshmikantham, V., Bainov, D.D., Simeonov, P.S.: Theory of Impulsive Differential Equations. World Scientific, Singapore (1989)
19. Liu, X., Zhang, S.H.: A cell population model described by impulsive PDE-s, existence and numerical approximation. Comp. Math. Appl. **36**(8), 1–11 (1998)
20. Midodashvili, B.: A nonlocal problem for fourth order hyperbolic equations with multiple characteristics. Electr. J. Differ. Equ. **2002**(85), 1–7 (2002)
21. Midodashvili, B.: Generalized Goursat problem for a spatial fourth order hyperbolic equation with dominated low terms. Proc. A. Razmadze Math. Inst. **138**, 43–54 (2005)
22. Nakhushev, A.M.: Problems with Shift for a Partial Differential Equations. Nauka, Moscow (2006). (in Russ.)
23. Perestyuk, N.A., Tkach, A.B.: Periodic solutions for weakly nonlinear partial system with pulse influense. Ukr. Math. J. **49**(4), 601–605 (1997)
24. Ptashnyck, B.I.: Incorrect Boundary Problems for Partial Differential Equations. Naukova dumka, Kiev (1984). (in Russ.)
25. Ptashnyck, B.Y., Il'kiv, V.S., Kmit', I.I., Polishuk, V.M.: Nonlocal Boundary Value Problems for Partial Differential Equations. Naukova dumka, Kiev (2002). (in Ukr.)
26. Rogovchenko, S.P.: Periodic solutions for hyperbolic impulsive systems. Preprint/Ukranian Academy of Sciences. No 88.3. Institute of Mathematics, Kiev (1988). (in Russian)
27. Samoilenko, A.M., Perestyuk, N.A.: Impulsive Differential Equations. World Scientific, Singapore (1995)
28. Tkach, A.B.: Numerical-analytic method of finding periodic solutions for systems of partial differential equations with pulse influence. Nonlinear Oscill. **4**(2), 278–288 (2001)

Application of Method of Differential Inequalities to Bounding the Rate of Convergence for a Class of Markov Chains

Anastasia Kryukova, Victoria Oshushkova, Alexander Zeifman, and Yacov Satin

Abstract We consider the linear system of differential equations $\frac{d\mathbf{p}}{dt} = A(t)\mathbf{p}$, which is the forward Kolmogorov system, for a class of Markov chains with 'batch' births and single deaths. We apply the method of differential inequalities for obtaining bounds on the rate of convergence for the system. A specific queueing model is considered and the corresponding limiting characteristics are computing.

Keywords Forward Kolmogorov system · Markov chains

1 Introduction and General Bounds

Let $\{X(t),\ t \geq 0\}$ be a continuous-time Markov chain with finite state space $\mathcal{X} = \{0, 1, \ldots, N\}$. Denote by $p_{ij}(s,t) = P\{X(t) = j\,|X(s) = i\}$, $i, j \geq 0,\ 0 \leq s \leq t$ the transition probabilities of $X(t)$ and by $p_i(t) = P\{X(t) = i\}$ – the probability that the Markov chain $X(t)$ is in state i at time t. Let $\mathbf{p}(t) = (p_0(t), p_1(t), \ldots)^T$ be the vector of state probabilities at the moment t.

Then the probabilistic dynamics of the process $\{X(t),\ t \geq 0\}$ is described by the forward Kolmogorov system

$$\frac{d\mathbf{p}}{dt} = A(t)\mathbf{p}, \tag{1}$$

where $A(t) = Q^T(t)$ is the transposed intensity matrix. All column sums of this matrix are zeros for any $t \geq 0$, and $A(t)$ is essentially nonnegative (i.e. all its off-diagonal elements are nonnegative for any $t \geq 0$).

We suppose that all 'intensity functions' $a_{ij}(t)$ are analytic in t for $t \geq 0$.

Consider a queueing model for a queue with batch arrivals and single services, see the first motivation in [2] and more recent studies in [1, 5, 6].

A. Kryukova · V. Oshushkova · Y. Satin
Vologda State University, Lenina, 15, Vologda, Russia

A. Zeifman (✉)
Institute of Informatics Problems FRC CSC RAS, Vologda Research Center RAS,
Vologda State University, Lenina, 15, Vologda, Russia
e-mail: a_zeifman@mail.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_8

Then we have $a_{ij}(t) = 0$ for $i < j - 1$, all arrival rates do not depend on the size of a queue, i.e. $a_{i+k,i}(t) = a_k(t)$ for $k \geq 1$, service rates $a_{i,i+1}(t) = \mu_{i+1}(t)$, and the matrix $A(t)$ has the following structure:

$$A(t) = \begin{pmatrix} a_{00}(t) & \mu_1(t) & 0 & 0 & \cdots & 0 & 0 \\ a_1(t) & a_{11}(t) & \mu_2(t) & 0 & \cdots & 0 & 0 \\ a_2(t) & a_1(t) & a_{22}(t) & \mu_3(t) & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ a_{N-1}(t) & a_{N-2}(t) & a_{N-3}(t) & a_{N-4}(t) & \cdots & a_{N-1,N-1}(t) & \mu_N(t) \\ a_N(t) & a_{N-1}(t) & a_{N-2}(t) & a_{N-3}(t) & \cdots & a_1(t) & a_{NN}(t) \end{pmatrix}. \tag{2}$$

Here we deal with a model of this class under additional suppositions $a_i(t) = 0,\ 1 \leq i \leq N-1,\ a_N(t) = a(t)$ (only arrival of all customers simultaneously is possible) and $\mu_i(t) \leq \mu_{i+1}(t)$ for any $i, t \geq 0$.

The difficulty of studying this model is due to the fact that it is not possible to apply the most convenient method of the logarithmic norm for it, see [5].

Now we get the following expression for the transposed intensity matrix:

$$A(t) = \begin{pmatrix} -a(t) & \mu_1(t) & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\mu_1(t) & \mu_2(t) & 0 & \cdots & 0 & 0 \\ 0 & 0 & -\mu_2(t) & \mu_3(t) & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & 0 & \cdots & -\mu_{N-1}(t) & \mu_N(t) \\ a(t) & 0 & 0 & 0 & \cdots & 0 & -\mu_N(t) \end{pmatrix}. \tag{3}$$

Put $p_0(t) = 1 - \sum\limits_{i \geq 1} p_i(t)$, then from (1) we obtain

$$\frac{d\mathbf{z}}{dt} = B(t)\,\mathbf{z} + \mathbf{f}(t), \tag{4}$$

where $\mathbf{f}(t) = (0, \ldots, 0, a(t))^T$, $\mathbf{z} = (p_1(t), p_2(t), \ldots, p_N(t))^T$,

$$B(t) = \begin{pmatrix} -\mu_1(t) & \mu_2(t) & 0 & \cdots & 0 & 0 \\ 0 & -\mu_2(t) & \mu_3(t) & \cdots & 0 & 0 \\ 0 & 0 & -\mu_3(t) & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & -\mu_{N-1}(t) & \mu_N(t) \\ -a(t) & -a(t) & -a(t) & \cdots & -a(t) & -\mu_N(t) - a(t) \end{pmatrix}. \tag{5}$$

All bounds on the rate of convergence to the limiting regime for $X(t)$ correspond to the same bounds of the solutions of system

$$\frac{d\mathbf{x}}{dt} = B(t)\mathbf{x}(t). \tag{6}$$

Denote by T upper triangular matrix

$$T = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \tag{7}$$

hence

$$T^{-1} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Let $\mathbf{u}(t) = T\mathbf{x}(t)$, then

$$\frac{d\mathbf{u}}{dt} = B^*(t)\mathbf{u}(t), \tag{8}$$

where $B^*(t) = TB(t)T^{-1}$, and

$$B^*(t) = \begin{pmatrix} -\mu_1(t) - a(t) & \mu_1(t) & 0 & 0 & \cdots & 0 & 0 \\ -a(t) & -\mu_2(t) & \mu_2(t) & 0 & \cdots & 0 & 0 \\ -a(t) & 0 & -\mu_3(t) & \mu_3(t) & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ -a(t) & 0 & 0 & 0 & \cdots & -\mu_{N-1}(t) & \mu_{N-1}(t) \\ -a(t) & 0 & 0 & 0 & \cdots & 0 & -\mu_N(t) \end{pmatrix}. \tag{9}$$

Once again, we note that the matrix $B^*(t)$ in is not essentially non-negative, and in such a situation the method of the logarithmic norm is inconvenient to apply (it gives poor results).

For the study of this system, we use the differential inequalities method, which was described in [3, 7].

Let d_i, $i = 1, \dots, N$ be nonzero numbers, and $\mathsf{D} = diag\,(d_1, d_2, \dots d_N)$ be a diagonal matrix:

$$\mathsf{D} = \begin{pmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ 0 & 0 & d_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_N \end{pmatrix}. \tag{10}$$

Put $\mathbf{w}(t) = \mathsf{D}\mathbf{u}(t)$, then we obtain from (8) the following system:

$$\frac{d\mathbf{w}}{dt} = B^{**}(t)\mathbf{w}(t), \tag{11}$$

where $B^{**}(t) = \mathsf{D}B^{*}(t)\mathsf{D}^{-1} =$

$$= \begin{pmatrix} -\mu_1(t) - a(t) & \mu_2(t)\cdot\frac{d_1}{d_2} & 0 & 0 & \cdots & 0 & 0 \\ -a(t)\cdot\frac{d_2}{d_1} & -\mu_2(t) & \mu_3(t)\cdot\frac{d_2}{d_3} & 0 & \cdots & 0 & 0 \\ -a(t)\cdot\frac{d_3}{d_1} & 0 & -\mu_3(t) & \mu_4(t)\cdot\frac{d_3}{d_4} & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ -a(t)\cdot\frac{d_{N-2}}{d_1} & 0 & 0 & 0 & \cdots & -\mu_{N-1}(t) & \mu_N(t)\cdot\frac{d_{N-1}}{d_N} \\ -a(t)\cdot\frac{d_{N-1}}{d_1} & 0 & 0 & 0 & \cdots & 0 & -\mu_N(t) \end{pmatrix}.$$

Let $\mathbf{u}(t)$ be an arbitrary solution of system (8).

Since the function $u_k(t)$ (the $k-$th coordinate of $\mathbf{u}(t)$) is analytic, it has a finite number of zeros on each interval. Consider an interval in which the signs of all the functions $u_k(t)$ do not change, say, (t_1, t_2). Choose the elements of the diagonal matrix such that signs of the entries d_i are equal with signs of corresponding coordinates $u_i(t)$ of the solution of system (8).

Since any $d_k u_k(t) > 0$ on the corresponding time interval, the sum $\sum_{k=1}^{N} d_k u_k = \|\mathbf{w}\|$ can be considered as the corresponding norm.

Moreover, we have the following inequalities:

$$\|\mathbf{u}\| \le N\|\mathbf{x}\|,\ \|\mathbf{x}\| \le 2\|\mathbf{u}\|,\ \|\mathbf{w}\| \le \max_k d_k\|\mathbf{u}\|,\ \|\mathbf{u}\| \le \left(\min_k d_k\right)^{-1}\|\mathbf{w}\|. \tag{12}$$

Let $B^{**}(t) = \left(b_{ij}^{**}(t)\right)_{i,j=1}^{N}$. Now, if the function $\alpha_D(t)$ is such that $\sum_{i=1}^{N} b_{ij}^{**}(t) \le -\alpha_D(t)$, $j = 1, ..., N$, then the following bound holds:

$$\frac{d\,\|\mathbf{w}\|}{dt} = \frac{d\left(\sum_{i=1}^{N} w_i\right)}{dt} = \sum_{j=1}^{N}\sum_{i=1}^{N} b_{ij}^{**}(t)w_j \le -\alpha_D(t)\,\|\mathbf{w}\|\,,$$

and hence

$$\|\mathbf{w}(t)\| \le e^{-\int_s^t \alpha_D(\tau)d\tau}\|\mathbf{w}(s)\|.$$

Therefore, in the original norm we get the following inequality

$$\|\mathbf{x}(t)\| \le \frac{2N\max_k d_k}{\min_k d_k} e^{-\int_s^t \alpha_D(\tau)d\tau}\|\mathbf{x}(s)\|, \tag{13}$$

for any $t_1 < s \le t < t_2$, and by continuity we get this inequality for $s = t_1$, $t = t_2$. Now we consider all such intervals (there is only a finite number 2^N of intervals with different sign combinations) and put $\alpha^*(t) = \min\{\alpha_D(t)\}$, $C = \max\left(\frac{\max_k d_k}{\min_k d_k}\right)$ where the minimum of $\alpha_D(t)$ and the maximum of $\frac{\max_k d_k}{\min_k d_k}$ is taken over all intervals with different sign combinations of coordinates of the solution. Finally, we obtain the following bound

$$\|\mathbf{x}(t)\| \le 2NCe^{-\int_0^t \alpha^*(\tau)d\tau}\|\mathbf{x}(0)\|. \tag{14}$$

In our case (in general, all intensities depend on the time t)

$$\sum_{i=1}^N w_i' = \left(-\mu_1 - a\cdot\left(1 + \frac{d_2}{d_1} + \frac{d_3}{d_1} + \cdots + \frac{d_N}{d_1}\right)\right)\cdot w_1 - \mu_2\cdot\left(1 - \frac{d_1}{d_2}\right)\cdot w_2$$

$$-\mu_3\cdot\left(1 - \frac{d_2}{d_3}\right)\cdot w_3 - \cdots - \mu_N\cdot\left(1 - \frac{d_{N-1}}{d_N}\right)\cdot w_N$$

(1) Let all $u_1, ..., u_N$ be positive. Since $\left(1 - \frac{d_i}{d_{i+1}}\right)$ must be positive, we have $d_{i+1} > d_i$. Suppose $d_1 := \varepsilon^N$, $d_2 := \varepsilon^{N-1}$, ..., $d_N := \varepsilon$, then

$$\sum_{i=1}^N w_i' = \left(-\mu_1 - a\cdot\left(1 + \frac{d_2}{d_1} + \frac{d_3}{d_1} + \cdots + \frac{d_N}{d_1}\right)\right)\cdot w_1 - \mu_2\cdot\left(1 - \frac{d_1}{d_2}\right)\cdot w_2$$

$$-\mu_3\cdot\left(1 - \frac{d_2}{d_3}\right)\cdot w_3 - \cdots - \mu_N\cdot\left(1 - \frac{d_{N-1}}{d_N}\right)\cdot w_N$$

$$= \left(-\mu_1 - a\cdot\left(1 + \frac{1}{\varepsilon} + \frac{1}{\varepsilon^2} + \cdots + \frac{1}{\varepsilon^{N-1}}\right)\right)\cdot w_1 - \mu_2\cdot(1-\varepsilon)\cdot w_2$$

$$-\mu_3\cdot(1-\varepsilon)\cdot w_3 - \cdots - \mu_N\cdot(1-\varepsilon)\cdot w_N,$$

and we have for the corresponding interval $\alpha_D = \min\{\mu_i\cdot(1-\varepsilon)\} = \mu_1\cdot(1-\varepsilon)$.

(2) Let all $u_1, ..., u_k$ be positive, and all $u_{k+1}, ..., u_N$ negative. Similarly $|d_{i+1}| > |d_i|$. Suppose $d_1 := \varepsilon^k$, $d_2 := \varepsilon^{k-1}$, ..., $d_k := \varepsilon$, $d_{k+1} := -\varepsilon^N$, $d_{k+2} := -\varepsilon^{N-1}$, ... , $d_N := -\varepsilon^{k+1}$, then

$$\sum_{i=1}^{N} w_i' = \left(-\mu_1 - a \cdot \left(1 + \frac{d_2}{d_1} + \frac{d_3}{d_1} + \cdots + \frac{d_N}{d_1}\right)\right) \cdot w_1 - \mu_2 \cdot \left(1 - \frac{d_1}{d_2}\right) \cdot w_2$$

$$-\mu_3 \cdot \left(1 - \frac{d_2}{d_3}\right) \cdot w_3 - \cdots - \mu_N \cdot \left(1 - \frac{d_{N-1}}{d_N}\right) \cdot w_N$$

$$= \left(-\mu_1 - a \cdot \left(1 + \frac{1}{\varepsilon} + \frac{1}{\varepsilon^2} + \cdots + \frac{1}{\varepsilon^{k-1}} - \varepsilon^{N-k} - \varepsilon^{N-k-1} - \ldots - \varepsilon\right)\right) \cdot w_1$$

$$-\mu_2 \cdot (1-\varepsilon) \cdot w_2 - \mu_3 \cdot (1-\varepsilon) \cdot w_3 - \cdots - \mu_k \cdot (1-\varepsilon) \cdot w_k - \mu_{k+1} \cdot \left(1 + \frac{1}{\varepsilon^{N-1}}\right) \cdot w_{k+1}$$

$$-\mu_{k+2} \cdot (1-\varepsilon) \cdot w_{k+2} - \ldots - \mu_N \cdot (1-\varepsilon) \cdot w_N.$$

In this case we also have the corresponding interval $\alpha_D = \min\{\mu_i \cdot (1-\varepsilon)\} = \mu_1 \cdot (1-\varepsilon)$.

Every time we changing sign on going from u_s to u_{s+1} we suppose $|d_{s+1}|$ be equal ε^m, where m is the number of the last element period of consistency.

Then we have $C = \varepsilon^{1-N}$, and the following bounds hold:

$$\|x(t)\| \leq 2N\varepsilon^{1-N} e^{-\mu_1 \cdot (1-\varepsilon)t} \|x(0)\|, \tag{15}$$

for the homogeneous Markov chain (constant intensities);
and

$$\|x(t)\| \leq 2N\varepsilon^{1-N} e^{-(1-\varepsilon)\int_0^t \mu_1(\tau)d\tau} \|x(0)\|, \tag{16}$$

in general situation.

2 Example

Consider here a specific queueing model with 1−periodic intensities. Let $a(t) = \lambda(t) = 2 + \sin(2\pi t)$ and $\mu_k(t) = k\,(2 + \cos(2\pi t))$. Then $A(t) =$

$$= \begin{pmatrix} -(2+\sin(2\pi t)) & 2+\cos(2\pi t) & 0 & \cdots & 0 & 0 \\ 0 & -(2+\cos(2\pi t)) & 2\cdot(2+\cos(2\pi t)) & \cdots & 0 & 0 \\ 0 & 0 & -2\cdot(2+\cos(2\pi t)) & \cdots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & -(N-1)(2+\cos(2\pi t)) & N\cdot(2+\cos(2\pi t)) \\ 2+\sin(2\pi t) & 0 & 0 & \cdots & 0 & -N\cdot(2+\cos(2\pi t)) \end{pmatrix},$$

$B^{**}(t) =$

$$\begin{pmatrix} -(2+\cos(2\pi t))-(2+\sin(2\pi t)) & 2\cdot(2+\cos(2\pi t))\cdot\frac{d_1}{d_2} & 0 & \cdots & 0 \\ -(2+\sin(2\pi t))\cdot\frac{d_2}{d_1} & -2\cdot(2+\cos(2\pi t)) & 3\cdot(2+\cos(2\pi t))\cdot\frac{d_2}{d_3} & \cdots & 0 \\ -(2+\sin(2\pi t))\cdot\frac{d_3}{d_1} & 0 & -3\cdot(2+\cos(2\pi t)) & \cdots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ -(2+\sin(2\pi t))\cdot\frac{d_{N-1}}{d_1} & 0 & 0 & \cdots & N\cdot(2+\cos(2\pi t))\cdot\frac{d_{N-1}}{d_N} \\ -(2+\sin(2\pi t))\cdot\frac{d_N}{d_1} & 0 & 0 & \cdots & -N\cdot(2+\cos(2\pi t)) \end{pmatrix},$$

and we have

$$\sum_{i=1}^{N} w_i' = (-(2+\cos(2\pi t)) - (2+\sin(2\pi t))\cdot(1+\frac{d_2}{d_1}+\frac{d_3}{d_1}+\cdots+\frac{d_N}{d_1}))\cdot w_1$$

$$-2\cdot(2+\cos(2\pi t))\cdot(1-\frac{d_1}{d_2})\cdot w_2 - 3\cdot(2+\cos(2\pi t))\cdot(1-\frac{d_2}{d_3})\cdot w_3$$

$$-\cdots-N\cdot(2+\cos(2\pi t))\cdot(1-\frac{d_{N-1}}{d_N})\cdot w_N.$$

Then we have the following bound on the rate of convergence

$$\|\mathbf{x}(t)\| \leq 2N\varepsilon^{1-N}e^{-(1-\varepsilon)t}\|\mathbf{x}(0)\|. \qquad (17)$$

Let $N = 200$. Then for any $\varepsilon \in (0, 1)$, we obtain the corresponding bound on the rate of convergence.

Denote by $E(t,k) = E(X(t)|X(0) = k)$ the conditional expected number of customers in the queue at instant t, provided that initially (at instant $t = 0$) k customers were present in the queue.

We compute here the probability of the empty queue $p_0(t)$ and the mathematical expectation of the number of customers in the queue $E(t,k)$, as it is shown on the Figs. 1, 2, 3 and 4.

These graphs are obtained using our standard approach (see detailed description in [4]) for solving numerically the forward Kolmogorov system on the corresponding interval and find approximately the limiting characteristics of this queueing model.

Note, that the bound (17) guarantees the coincidence of the probability characteristics for the queue-length process with different initial conditions with a predetermined accuracy for the corresponding (sufficiently large) values of t. In fact, as the graphs show, the difference is already quite small at $t \geq 17$.

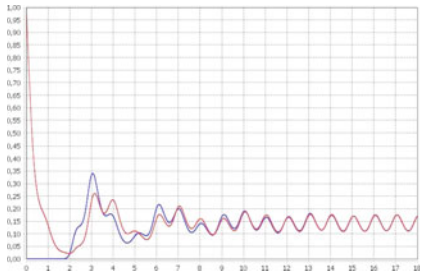


Fig. 1 Example. Probability of the empty queue for $t \in [0, 18]$ with initial conditions $X(0) = 0$ (red) and $X(0) = 200$ (blue)

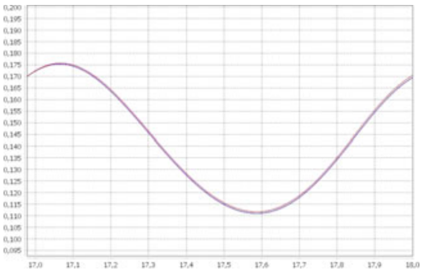


Fig. 2 Example. Probability of the empty queue for $t \in [17, 18]$ with initial conditions $X(0) = 0$ and $X(0) = 200$

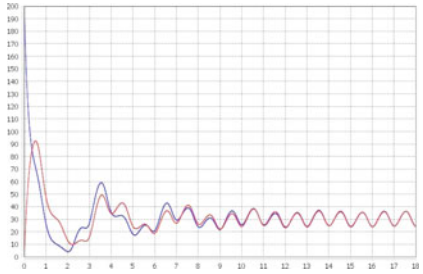


Fig. 3 Example. The mean $E(t, k)$ for $t \in [0, 18]$ with initial conditions $X(0) = 0$ (red) and $X(0) = 200$ (blue)

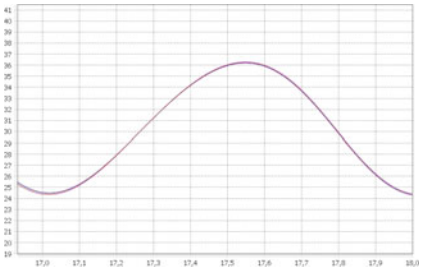


Fig. 4 Example. The mean $E(t, k)$ for $t \in [17, 18]$ with initial conditions $X(0) = 0$ and $X(0) = 200$

Acknowledgements This research was supported by Russian Science Foundation under grant 19-11-00020.

References

1. Li, J., Zhang, L.: M X/M/c queue with catastrophes and state-dependent control at idle time. Front. Math. China **12**(6), 1427–1439 (2017)
2. Nelson, R., Towsley, D., Tantawi, A.N.: Performance analysis of parallel processing systems. IEEE Trans. Softw. Eng. **14**(4), 532–540 (1988)
3. Satin, Y., Zeifman, A., Kryukova, A.: On the rate of convergence and limiting characteristics for a nonstationary queueing model. Mathematics **7**(8), 678 (2019)
4. Zeifman, A., Satin, Y., Korolev, V., Shorgin, S.: On truncations for weakly ergodic inhomogeneous birth and death processes. Int. J. Appl. Math. Comput. Sci. **24**, 503–518 (2014)
5. Zeifman, A., Razumchik, R., Satin, Y., Kiseleva, K., Korotysheva, A., Korolev, V.: Bounds on the rate of convergence for one class of inhomogeneous Markovian queueing models with possible batch arrivals and services. Int. J. Appl. Math. Comput. Sci. **28**, 141–154 (2018)
6. Zeifman, A., Sipin, A., Korolev, V., Shilova, G., Kiseleva, K., Korotysheva, A., Satin, Y.: On sharp bounds on the rate of convergence for finite continuous-time Markovian queueing models. LNCS, vol. 10672, pp. 20–28 (2018)
7. Zeifman, A., Satin, Y., Kiseleva, K., Kryukova, A.: Applications of differential inequalities to bounding the rate of convergence for continuous-time Markov chains. In: AIP Conference Proceedings, vol. 2116, p. 090009 (2019)

The Method of Fractional Steps for the Numerical Solution of a Multidimensional Heat Conduction Equation with Delay for the Case of Variable Coefficient of Heat Conductivity

Andrei Lekomtsev

Abstract Multidimensional parabolic equations with delay effects in the time component for the case of variable coefficient of heat conductivity depending on spatial and temporal variables are considered. The method of fractional steps is constructed for the numerical solution of these equations. The order of approximation error for the constructed method, stability, and order of convergence are investigated. A theorem is obtained on the order of convergence of the method of fractional steps, which uses the methods from the general theory of difference schemes and the technique of the investigation of difference schemes for solving functional differential equations. Results of calculating test example with variable concentrated and distributed time delay are presented.

Keywords Numerical solution · Multidimensional heat conduction equation · Heat conductivity · Stability · Order of convergence

1 Introduction

The work is devoted to the development and study of the convergence of numerical algorithms for solving multidimensional parabolic equations with a delay effect, constant, variable, or distributed. Such effects are found in many mathematical models [13, 14] and numerical methods for solving them were studied in a number of works, see for example [1–3, 11, 12, 15].

A number of papers [12, 15] are devoted to the application of the method of lines, at which discretization is performed only in spatial variable. However, with such discretization, the problem of stiffness arises. In addition, this method is difficult to apply for multidimensional in space equations.

In some works, see [3] and the bibliography in this work, grid methods were investigated from a general point of view. The main idea in this approach consists in the introduction of the intermediate interpolation space. This raises the problem

A. Lekomtsev (✉)
Ural Federal University, Yekaterinburg, Russia
e-mail: avlekomtsev@urfu.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_9

of solving systems of large dimension, as well as the investigation of the stability of these algorithms.

In the paper [7] this approach was complemented by the idea of separating the finite-dimensional current state of the system and the infinite-dimensional (functional) prehistory. The finite-dimensional component is used to construct complete analogs of numerical methods known for equations without delay, and simple methods of interpolation of discrete prehistory with given properties are used to take into account functional delay. To implement implicit methods, extrapolation of discrete prehistory is used. This approach allows us to develop effective algorithms that can serve as the basis for creating application packages for solving such problems.

In the framework of this approach, in this paper, we consider the multidimensional parabolic equation with variable coefficients of heat conductivity and with a delay effect included only in the inhomogeneous term. Used algorithms linear part, known for parabolic equations (fractional step method). Earlier in the work [4] was considered two-dimensional version of this method for the equation with constant coefficients of heat conductivity. In the work [5] was considered one-dimensional version with variable coefficient of heat conductivity.

We briefly review the content of the work.

After the statement of the problem, the main assumptions about the equation are made, which are needed later for proof of the theorem of convergence. Discretization of the problem is carried out. With the help of the interpolation constructs the analog of the fractional steps method is constructed. The algorithm is reduced to solving linear algebraic systems of a tridiagonal structure. The order of the local error (residual) of the algorithm is investigated.

The main result of the work is to prove the convergence of this algorithm. The convergence of these algorithms is justified by applying the general theory of difference schemes [8], and the general methodology of research of difference schemes for solving equations with heredity [6]. The latter method uses the ideas of the work [10], developed for ordinary differential equations without delay. However, the presence of variable coefficients of heat conductivity required modification of the general difference scheme. As part of this modification, the basic concepts of temporary grid are introduced; discrete model at each moment of time as an element of finite-dimensional normed space, prehistory of discrete model, interpolation space, explicit step-by-step formula, initial values, function of exact values. In the step-by-step formula, the operator of transition is highlighted, its properties determine the stability of the scheme. The main statement is given - the theorem on the order of convergence of the scheme, which depends on the residual order with interpolation.

Next, the constructed algorithm is embedded in this scheme. The relation between the residual order without interpolation, the order of interpolation, and the residual order with interpolation is established. The condition that guarantees the stability of the developed scheme is investigated in detail. Using the embedding of the algorithm in the general scheme, we obtain the theorem on the orders of convergence, the order of convergence of the methods is linear in time step and quadratic in spatial steps.

The paper finishes with a presentation of the results of numerical experiments. A test example with variable concentrated and distributed delays for the case of variable coefficient of heat conductivity in the case of dependence on spatial and temporal variables is given.

2 Problem Statement and Main Assumptions

Consider the following p-dimensional heat conduction equation with after effect:

$$\frac{\partial u}{\partial t} = \sum_{\beta=1}^{p} \frac{\partial}{\partial x_\beta}(K_\beta(x,t)\frac{\partial u}{\partial x_\beta}) + f(x,t,u(x,t),u_t(x,\cdot)), \tag{1}$$

where $x = (x_1, \dots, x_p) \in \overline{G} = \{0 \le x_\beta \le X_\beta, \quad \beta = \overline{1,p}\}$—space variable and $t \in [t_0, \theta] \subset \mathbb{R}^1$—time variable; $u(x,t) \in \mathbb{R}^1$ is the required function; $u_t(x,\cdot) = \{u(x,t+s), \quad -\tau \le s < 0\}$ is the prehistory of the required function by the time t; and τ is the value of delay. Let the initial and boundary conditions be given.

$$u(x,t) = \varphi(x,t),\ x \in \overline{G},\ t \in [t_0 - \tau, t_0], \tag{2}$$

$$u|_\Gamma = 0,\ \Gamma - \text{boundary of } G. \tag{3}$$

We assume that coefficients $K_\beta(x,t)$ satisfies the following conditions:

$$0 < c_1 \le K_\beta(x,t) \le c_2,\ x \in \overline{G},\ t \in [t_0, \theta], \tag{4}$$

$$\frac{|K_\beta(x,t) - K_\beta(x,t-\Delta)|}{\Delta} \le c_3 K_\beta(x,t-\Delta),\ x \in \overline{G},\ t > t_0.\ [8] \tag{5}$$

We will also assume that the functions $K_\beta(x,t),\ \varphi(x,t)$ and functional f have the required properties for the existence of a unique solution $u(x,t)$ of the problem (1)–(3), and the solution $u(x,t)$ is understood in the classical sense. We assume that the function $u(x,t)$ has a certain degree of smoothness, which will be indicated in the further reasonings.

We denote by $Q = Q[-\tau, 0]$ the set of functions $q(s)$ that are piecewise continuous on the interval $[-\tau, 0]$ with a finite number of points of discontinuity of the first kind and right continuous at the points of discontinuity. We define the norm of a function on $Q[-\tau, 0]$ by the relation $\|q(\cdot)\|_{Q[-\tau,0]} = \max\limits_{-\tau \le s \le 0} |q(s)|$.

$Q[-\tau, 0)$ is the reduction of the space $Q[-\tau, 0]$ on the half-open interval $[-\tau, 0)$, such that the functions $q(s) \in Q[-\tau, 0)$ have a finite left-hand limit at zero.

Suppose that the functional $f(x,t,u,u(\cdot))$ is Lipschitz [5] with constant L_f with respect to the last two arguments.

3 The Method of Fractional Steps

Without loss of generality, we assume that $p = 3$.

We divide the intervals $[0, X_1]$, $[0, X_2]$, $[0, X_3]$ into parts with steps $h_1 = X_1/N_1$, $h_2 = X_2/N_2$, $h_3 = X_3/N_3$ respectively. Let us introduce a grid $\overline{\omega}_h$. $\overline{\omega}_h$ is the set of points (nodes) $x^{i_1i_2i_3} = (x_1^{i_1}, x_2^{i_2}, x_3^{i_3})$, where $x_1^{i_1} = i_1h_1,\ i_1 = \overline{0, N_1}; x_2^{i_2} = i_2h_2,\ i_2 = \overline{0, N_2}; x_3^{i_3} = i_3h_3,\ i_3 = \overline{0, N_3}$. That is

$$\overline{\omega}_h = \{x^{i_1i_2i_3},\ i_1 = \overline{0, N_1},\ i_2 = \overline{0, N_2},\ i_3 = \overline{0, N_3}\},$$

$$\omega_h = \{x^{i_1i_2i_3},\ i_1 = \overline{1, N_1 - 1},\ i_2 = \overline{1, N_2 - 1},\ i_3 = \overline{1, N_3 - 1}\},\ \gamma_h = \overline{\omega}_h \backslash \omega_h.$$

We divide the interval $[t_0, \theta]$ into parts with step $\Delta > 0$, introducing the points $t_k = t_0 + k\Delta,\ k = \overline{0, M}$. We denote by $t_{k+\frac{1}{2}} = t_0 + (k + 1/2)\Delta,\ k = \overline{0, M - 1}$. We assume that the value $\tau/\Delta = m$ is an integer. We denote by $u_k^{i_1i_2i_3}$ approximations of the exact solution $u(x^{i_1i_2i_3}, t_k)$ at the node $(x^{i_1i_2i_3}, t_k)$. For every fixed $i_1 = \overline{0, N_1}$, $i_2 = \overline{0, N_2}$, $i_3 = \overline{0, N_3}$ we introduce the discrete prehistory by the time t_k, $k = \overline{0, M}$: $\{u_l^{i_1i_2i_3}\}_k = \{u_l^{i_1i_2i_3}, k - m \le l \le k\}$.

Definition 1. The operator of interpolation-extrapolation of the discrete prehistory is the mapping $I : \{u_l^{i_1i_2i_3}\}_k \to v_k^{i_1i_2i_3}(\cdot) \in Q[-\tau, \Delta]$.

Definition 2. The operator of interpolation-extrapolation I has order of error Δ^{p_0} on the exact solution if there exist constants C_1 and C_2 such that, for all $i_1, i_2,\ i_3,\ k$ and $t \in [t_k - \tau, t_{k+1}]$ the following inequality holds:

$$\|v_k^{i_1i_2i_3}(t) - u(x^{i_1i_2i_3}, t)\|_{Q[-\tau,\Delta]} \le C_1 \max_{k-m\le l\le k} |u_l^{i_1i_2i_3} - u(x^{i_1i_2i_3}, t_l)| + C_2\Delta^{p_0}.$$

Then, for example, the piecewise constant interpolation with extrapolation by extension has order of error Δ.

$$v_k^{i_1i_2i_3}(t) = \begin{cases} \varphi(x^{i_1i_2i_3}, t),\ t \in [t_0 - \tau, t_0], \\ u_{l-1}^{i_1i_2i_3},\ t \in [t_{l-1}, t_l],\ 1 \le l \le k, \\ u_k^{i_1i_2i_3},\ t \in [t_k, t_{k+1}]. \end{cases}$$

We use the notations

$$\Lambda_1(t)u_k^{i_1i_2i_3} = \frac{1}{h_1}[K_1(x^{i_1+\frac{1}{2}i_2i_3}, t)\frac{u_k^{i_1+1i_2i_3} - u_k^{i_1i_2i_3}}{h_1} - K_1(x^{i_1-\frac{1}{2}i_2i_3}, t)\frac{u_k^{i_1i_2i_3} - u_k^{i_1-1i_2i_3}}{h_1}],$$

$$\Lambda_2(t)u_k^{i_1i_2i_3} = \frac{1}{h_2}[K_2(x^{i_1i_2+\frac{1}{2}i_3}, t)\frac{u_k^{i_1i_2+1i_3} - u_k^{i_1i_2i_3}}{h_2} - K_2(x^{i_1i_2-\frac{1}{2}i_3}, t)\frac{u_k^{i_1i_2i_3} - u_k^{i_1i_2-1i_3}}{h_2}],$$

$$\Lambda_3(t)u_k^{i_1i_2i_3} = \frac{1}{h_3}[K_3(x^{i_1i_2i_3+\frac{1}{2}}, t)\frac{u_k^{i_1i_2i_3+1} - u_k^{i_1i_2i_3}}{h_3} - K_3(x^{i_1i_2i_3-\frac{1}{2}}, t)\frac{u_k^{i_1i_2i_3} - u_k^{i_1i_2i_3-1}}{h_3}],$$

where $x^{i_1\pm\frac{1}{2}i_2i_3} = (x_1^{i_1} \pm \frac{h_1}{2}, x_2^{i_2}, x_3^{i_3})$, $x^{i_1i_2\pm\frac{1}{2}i_3} = (x_1^{i_1}, x_2^{i_2} \pm \frac{h_2}{2}, x_3^{i_3})$, $x^{i_1i_2i_3\pm\frac{1}{2}} = (x_1^{i_1}, x_2^{i_2}, x_3^{i_3} \pm \frac{h_3}{2})$. For the transition from layer k to layer $k+1$ introduce auxiliary layers $k+1/3$ and $k+2/3$. Then for $0 \le s \le 1$ we consider a family of methods

$$\frac{u_{k+\frac{1}{3}}^{i_1i_2i_3} - u_k^{i_1i_2i_3}}{\Delta} = s\Lambda_1(t_{k+\frac{1}{2}})u_{k+\frac{1}{3}}^{i_1i_2i_3} + (1-s)\Lambda_1(t_{k+\frac{1}{2}})u_k^{i_1i_2i_3} + F_k^{i_1i_2i_3}\big(v_k^{i_1i_2i_3}(\cdot)\big), \quad (6)$$

$$\frac{u_{k+\frac{2}{3}}^{i_1i_2i_3} - u_{k+\frac{1}{3}}^{i_1i_2i_3}}{\Delta} = s\Lambda_2(t_{k+\frac{1}{2}})u_{k+\frac{2}{3}}^{i_1i_2i_3} + (1-s)\Lambda_2(t_{k+\frac{1}{2}})u_{k+\frac{1}{3}}^{i_1i_2i_3} + F_k^{i_1i_2i_3}\big(v_k^{i_1i_2i_3}(\cdot)\big), \quad (7)$$

$$\frac{u_{k+1}^{i_1i_2i_3} - u_{k+\frac{2}{3}}^{i_1i_2i_3}}{\Delta} = s\Lambda_3(t_{k+\frac{1}{2}})u_{k+1}^{i_1i_2i_3} + (1-s)\Lambda_3(t_{k+\frac{1}{2}})u_{k+\frac{2}{3}}^{i_1i_2i_3} + F_k^{i_1i_2i_3}\big(v_k^{i_1i_2i_3}(\cdot)\big), \quad (8)$$

where $i_1 = \overline{1, N_1 - 1}$, $i_2 = \overline{1, N_2 - 1}$, $i_3 = \overline{1, N_3 - 1}$, $k = \overline{0, M-1}$, with the initial and boundary conditions

$$u_0^{i_1i_2i_3} = \varphi(x^{i_1i_2i_3}, t_0),\ i_1 = \overline{0, N_1},\ i_2 = \overline{0, N_2},\ i_3 = \overline{0, N_3}, \quad (9)$$

$$v_0^{i_1i_2i_3}(t) = \varphi(x^{i_1i_2i_3}, t),\ t < t_0,\ i_1 = \overline{0, N_1},\ i_2 = \overline{0, N_2},\ i_3 = \overline{0, N_3}, \quad (10)$$

$$u_k^{i_1i_2i_3}|_\Gamma = 0,\ u_{k+\frac{1}{3}}^{i_1i_2i_3}|_\Gamma = 0,\ u_{k+\frac{2}{3}}^{i_1i_2i_3}|_\Gamma = 0,\ u_{k+1}^{i_1i_2i_3}|_\Gamma = 0,\ k = \overline{0, M-1}. \quad (11)$$

As a functional $F_k^{i_1i_2i_3}\big(v_k^{i_1i_2i_3}(\cdot)\big)$, we will consider $\dfrac{1}{3}f(x^{i_1i_2i_3}, t_{k+\frac{1}{2}}, u_{k+\frac{1}{2}}^{i_1i_2i_3}, v_{k+\frac{1}{2}}^{i_1i_2i_3}(\cdot))$. Note that due to interpolation and extrapolation, the value of the functional $F_k^{i_1i_2i_3}\big(v_k^{i_1i_2i_3}(\cdot)\big)$ is explicitly calculated. If $s = 0$, then we get an explicit scheme. If $0 < s \le 1$, then system (6)–(8) for any fixed k is a chain from three linear tridiagonal systems of the equations with respect to $u_{k+\frac{1}{3}}^{i_1i_2i_3}$, $u_{k+\frac{2}{3}}^{i_1i_2i_3}$ and $u_{k+1}^{i_1i_2i_3}$ respectively with diagonal dominances, which are effectively solved by the sweep method.

We introduce the space $H = \overset{\circ}{\Omega}_h$ as the set of grid functions defined on $\overline{\omega}_h$ and equal to zero on γ_h. The scalar product and norm in $\overset{\circ}{\Omega}_h$ is defined as follows [8]:

$$(y, z) = \sum_{x\in\omega_h} y(x)z(x)h_1h_2h_3,\ \|y\| = \sqrt{(y, y)}. \quad (12)$$

We use the notations

$$B_\beta(t) = E - s\Delta\Lambda_\beta(t),\ \beta = 1, 2, 3. \quad (13)$$

By virtue the condition (4) the operators—$\Lambda_\beta(t)$ are positive and self-adjoint [8] for all t in the sense of the scalar product (12). Hence it is clear that the operators $B_\beta(t)$ are also positive, self-adjoint, and inverse operators $B_\beta^{-1}(t)$ exists by virtue of their positivity. We denote

$$S_\beta(t) = E + \Delta B_\beta^{-1}(t)\Lambda_\beta(t),\ \ \beta = 1, 2, 3. \tag{14}$$

Then the system (6)–(8) is converted to the following form

$$\begin{aligned}
u_{k+\frac{1}{3}}^{i_1 i_2 i_3} &= S_1(t_{k+\frac{1}{2}})u_k^{i_1 i_2 i_3} + \Delta B_1^{-1}(t_{k+\frac{1}{2}})F_k^{i_1 i_2 i_3}\big(v_k^{i_1 i_2 i_3}(\cdot)\big),\\
u_{k+\frac{2}{3}}^{i_1 i_2 i_3} &= S_2(t_{k+\frac{1}{2}})u_{k+\frac{1}{3}}^{i_1 i_2 i_3} + \Delta B_2^{-1}(t_{k+\frac{1}{2}} F_k^{i_1 i_2 i_3}\big(v_k^{i_1 i_2 i_3}(\cdot)\big),\\
u_{k+1}^{i_1 i_2 i_3} &= S_3(t_{k+\frac{1}{2}})u_{k+\frac{2}{3}}^{i_1 i_2 i_3} + \Delta B_3^{-1}(t_{k+\frac{1}{2}})F_k^{i_1 i_2 i_3}\big(v_k^{i_1 i_2 i_3}(\cdot)\big).
\end{aligned}$$

We use the notation $\bar{t} = t_{k+\frac{1}{2}}$. Hence

$$\begin{aligned}
u_{k+1}^{i_1 i_2 i_3} = \ & S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})u_k^{i_1 i_2 i_3} + \Delta[S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t}) + S_3(\bar{t})B_2^{-1}(\bar{t})\\
& + B_3^{-1}(\bar{t})]F_k^{i_1 i_2 i_3}\big(v_k^{i_1 i_2 i_3}(\cdot)\big),
\end{aligned}$$

where $i_1 = \overline{1, N_1 - 1},\ i_2 = \overline{1, N_2 - 1},\ i_3 = \overline{1, N_3 - 1},\ k = \overline{0, M - 1}$. It is known [8] that for all t the following inequality holds

$$\| - \Lambda_\beta(t)\| \geq \frac{8c_1}{X_\beta^2} > 0,\ \ \beta = 1, 2, 3. \tag{15}$$

In view of (13), (15) for all t we obtain the estimates

$$\|B_\beta^{-1}(t)\| \leq \frac{1}{1 + \frac{8s\Delta C_1}{X_\beta^2}} < 1,\ \ \beta = 1, 2, 3. \tag{16}$$

In Sect. 5 we will show that for all t the following inequality holds

$$\|S_\beta(t)\| \leq 1,\ \ \beta = 1, 2, 3.$$

Definition 3. The residual of the method of fractional steps (6)–(11) is called

$$\begin{aligned}
\Psi_k^{i_1 i_2 i_3} = & \frac{u(x^{i_1 i_2 i_3}, t_{k+1}) - S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})u(x^{i_1 i_2 i_3}, t_k)}{\Delta} - [S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t})\\
& + S_3(\bar{t})B_2^{-1}(\bar{t}) + B_3^{-1}(\bar{t})]F_k^{i_1 i_2 i_3}\big(u_{t_k}(x^{i_1 i_2 i_3}, \cdot)\big).
\end{aligned} \tag{17}$$

Determining the order of the residual of the method (6)–(11) is performed using the Taylor expansion of the function $u(x, t)$ (under conditions of corresponding smoothness).

Theorem 1. *We will assume that the coefficients of heat conductivity $K_1(x, t)$, $K_2(x, t)$, $K_3(x, t)$ are three times continuously differentiable in x. We will also assume that the exact solution $u(x, t)$ of problem (1)–(3) is four times continuously differentiable in x and three times continuously differentiable in t, and the second* *derivatives of the solution $u(x,t)$ with respect to x are continuously differentiable in t. Then, the residual of the method (6)–(11) has order $\Delta + h_1^2 + h_2^2 + h_3^2$ for any s.*

Proof. In a neighborhood of the point $(x^{i_1 i_2 i_3}, \bar{t})$ we expand the exact solution $u(x,t)$ of problem (1)–(3) in Taylor's series. In view of condition (4) and under the assumptions of the theorem the following equality holds [8]

$$\sum_{\beta=1}^{3} \frac{\partial}{\partial x_\beta}(K_\beta(x,\bar{t})\frac{\partial u}{\partial x_\beta})(x^{i_1 i_2 i_3}, \bar{t})$$
$$= (\Lambda_1(\bar{t}) + \Lambda_2(\bar{t}) + \Lambda_3(\bar{t}))u(x^{i_1 i_2 i_3}, \bar{t}) + O(h_1^2 + h_2^2 + h_3^2). \tag{18}$$

We transform the operator $S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})$ in the definition of the residual (17) according to the notations (14)

$$\begin{aligned} S_3(\bar{t})S_2(\bar{t})S_1(\bar{t}) &= (E + \Delta B_3^{-1}(\bar{t})\Lambda_3(\bar{t}))(E + \Delta B_2^{-1}(\bar{t})\Lambda_2(\bar{t}))(E + \Delta B_1^{-1}(\bar{t})\Lambda_1(\bar{t})) \\ &= E + \Delta[B_1^{-1}(\bar{t})\Lambda_1(\bar{t}) + B_2^{-1}(\bar{t})\Lambda_2(\bar{t}) + B_3^{-1}(\bar{t})\Lambda_3(\bar{t})] \\ &+\Delta^2[B_3^{-1}(\bar{t})\Lambda_3(\bar{t})B_2^{-1}(\bar{t})\Lambda_2(\bar{t}) + B_3^{-1}(\bar{t})\Lambda_3(\bar{t})B_1^{-1}(\bar{t})\Lambda_1(\bar{t}) \\ &+B_2^{-1}(\bar{t})\Lambda_2(\bar{t})B_1^{-1}(\bar{t})\Lambda_1(\bar{t})] + \Delta^3 B_3^{-1}(\bar{t})\Lambda_3(\bar{t})B_2^{-1}(\bar{t})\Lambda_2(\bar{t})B_1^{-1}(\bar{t})\Lambda_1(\bar{t}). \end{aligned} \tag{19}$$

We transform the operators $S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t})$, $S_3(\bar{t})B_2^{-1}(\bar{t})$ in the definition (17)

$$S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t}) = (E + \Delta B_3^{-1}(\bar{t})\Lambda_3(\bar{t}))(E + \Delta B_2^{-1}(\bar{t})\Lambda_2(\bar{t}))B_1^{-1}$$
$$= (E + \Delta(B_3^{-1}(\bar{t})\Lambda_3(\bar{t}) + B_2^{-1}(\bar{t})\Lambda_2(\bar{t})) + \Delta^2 B_3^{-1}(\bar{t})\Lambda_3(\bar{t})B_2^{-1}(\bar{t})\Lambda_2(\bar{t}))B_1^{-1}. \tag{20}$$
$$S_3(\bar{t})B_2^{-1}(\bar{t}) = (E + \Delta B_3^{-1}(\bar{t})\Lambda_3(\bar{t}))B_2^{-1}(\bar{t}). \tag{21}$$

From (17), (18)–(21), positive and boundedness of operators $B_\beta^{-1}(\bar{t})$ we have

$$\begin{aligned} \Psi_k^{i_1 i_2 i_3} &= \frac{u(x^{i_1 i_2 i_3}, t_{k+1}) - u(x^{i_1 i_2 i_3}, t_k)}{\Delta} - (\Lambda_1(\bar{t}) + \Lambda_2(\bar{t}) + \Lambda_3(\bar{t}))u(x^{i_1 i_2 i_3}, t_k) \\ &-3F_k^{i_1 i_2 i_3}\big(u_{t_k}(x^{i_1 i_2 i_3}, \cdot)\big) + O(\Delta + h_1^2 + h_2^2 + h_3^2) = \frac{\partial u}{\partial t}(x^{i_1 i_2 i_3}, \bar{t}) + O(\Delta^2) \\ &-(\Lambda_1(\bar{t}) + \Lambda_2(\bar{t}) + \Lambda_3(\bar{t}))u(x^{i_1 i_2 i_3}, \bar{t}) + O(\Delta) - 3F_k^{i_1 i_2 i_3}\big(u_{t_k}(x^{i_1 i_2 i_3}, \cdot)\big) \\ &+O(\Delta + h_1^2 + h_2^2 + h_3^2) = \frac{\partial u}{\partial t}(x^{i_1 i_2 i_3}, \bar{t}) - \sum_{\beta=1}^{3} \frac{\partial}{\partial x_\beta}(K_\beta(x,\bar{t})\frac{\partial u}{\partial x_\beta})(x^{i_1 i_2 i_3}, \bar{t}) \\ &-f(x^{i_1 i_2 i_3}, \bar{t}, u(x^{i_1 i_2 i_3}, \bar{t}, u_{\bar{t}}(x^{i_1 i_2 i_3}, \cdot))) + O(\Delta + h_1^2 + h_2^2 + h_3^2). \end{aligned}$$

From the fact that $u(x,t)$ is the solution of the equation (1), it follows

$$|\Psi_k^{i_1 i_2 i_3}| \le C_\Psi(\Delta + h_1^2 + h_2^2 + h_3^2).$$

The conclusion of the theorem follows from this relation. □

4 General Difference Scheme with Aftereffect and Its Order of Convergence

We have a segment $[t_0, \theta]$ and a number $\tau > 0$—the value of the delay. A step of a grid is defined as a number $\Delta > 0$, such that $\tau/\Delta = m$ is an integer. $\{\Delta\}$—the set of steps. A uniform grid is defined as a finite set of numbers

$$\Sigma_\Delta = \{t_k = t_0 + k\Delta \in [t_0 - \tau, \theta], k = -m, \dots, M\}.$$

We denote $\Sigma_\Delta^- = \{t_k \in \Sigma_\Delta, k \leq 0\}$, $\Sigma_\Delta^+ = \{t_k \in \Sigma_\Delta, k \geq 0\}$. A discrete model is, by definition, a grid function $t_k \in \Sigma_\Delta \to y(t_k) = y_k \in Y,\ k = -m, \dots, M$, where Y is a q - dimensional normed space with norm $\|\cdot\|_Y$. We assume that the dimension q depends on a number $h > 0$.

For $k \geq 0$ the prehistory of the discrete model by the time t_k is, by definition, the set $\{y_l\}_k = \{y_l \in Y,\ l = k - m, \dots, k\}$. Let V (an interpolation space) be a linear normed space with norm $\|\cdot\|_V$. A mapping $I : I(\{y_l\}_k) = v \in V$—operator of the interpolation of the discrete prehistory of the model.

We assume that the interpolation operator satisfies the Lipschitz condition [5] with constant L_I. Starting values of the model are, by definition, the function $\Sigma_\Delta^- \to Y$:

$$y(t_k) = y_k,\ k = -m, \dots, 0. \tag{22}$$

The formula of the advance of the model by a step is defined by the algorithm

$$y_{k+1} = S_k y_k + \Delta\Phi(t_k, I(\{y_l\}_k), \Delta), \tag{23}$$

here $\Phi : \Sigma_\Delta^+ \times V \times \{\Delta\} \to Y$—the function of advance by a step; $S_k : Y \to Y$ is a linear transition operator.

Thus, a discrete model is defined by starting values (22), an interpolation operator and formula of advance by a step (23). We will suppose that the function $\Phi(t_k, v, \Delta)$ in (23) is Lipschitz [5] with constant L_Φ with respect to the second argument. The function of exact values is defined by the mapping

$$Z(t_k, \Delta) = z_k \in Y,\ k = -m, \dots, M.$$

We assume that the specification of the function of exact values is a consequence of the specification of an exact solution to the problem (1)–(3). We will say that starting values of the model have order $\Delta^{p_1} + h^{p_2}$ if there exists a constant C such that

$$\|z_k - y_k\|_Y \leq C(\Delta^{p_1} + h^{p_2}),\ k = -m, \dots, 0.$$

We will say that the method (23) converges with order $\Delta^{p_1} + h^{p_2}$ if there exists a constant C such that

$$\|z_k - y_k\|_Y \leq C(\Delta^{p_1} + h^{p_2}),\ k = -m, \ldots, M.$$

Method (23) is called stable if for all $k = 0, \ldots, M-1$ is satisfied the inequality $\|S_k\|_Y \leq 1$, where

$$\|S_k\|_Y = \sup_{y \neq 0} \frac{\|S_k y\|_Y}{\|y\|_Y},\ k = 0, \ldots, M-1. \tag{24}$$

An approximation error with interpolation (residual) is called the following grid function

$$d_k = (z_{k+1} - S_k z_k)/\Delta - \Phi(t_k, I(\{z_l\}_k), \Delta),\ k = 0, \ldots, M-1. \tag{25}$$

We will say that method (23) has the order of approximation error with interpolation $\Delta^{p_1} + h^{p_2}$ if there exists a constant C such that

$$\|d_k\|_Y \leq C(\Delta^{p_1} + h^{p_2}),\ k = 1, \ldots, M.$$

Theorem 2. *Suppose that the interpolation operator I satisfies the Lipschitz condition and the function Φ satisfies the Lipschitz condition with respect to the second argument. Suppose also that the starting values have order $\Delta^{p_1} + h^{p_2}$, where $p_1 > 0$ and $p_2 > 0$. Let the error of approximation with interpolation has order $\Delta^{p_3} + h^{p_4}$, where $p_3 > 0$ and $p_4 > 0$. Suppose also that the method (23) is stable. Then, the method (23) converges. Also, the order of the convergence of the method (23) is at least $\Delta^{\min\{p_1,p_3\}} + h^{\min\{p_2,p_4\}}$.*

Proof. We use the notation $\delta_k = z_k - y_k$, $k = -m, \ldots, M$. Then we have

$$\delta_{k+1} = S_k\delta_k + \Delta\widehat{\delta}_k + \Delta d_k,\ k = 0, \ldots, M-1,\ \text{where} \tag{26}$$

$$\widehat{\delta}_k = \Phi(t_k, I(\{z_l\}_k), \Delta) - \Phi(t_k, I(\{y_l\}_k), \Delta).$$

The assumptions that the mappings Φ and I are Lipschitz imply that

$$\|\widehat{\delta}_k\|_Y \leq K \max_{k-m \leq l \leq k} \{\|\delta_l\|_Y\},\ K = L_\Phi L_I. \tag{27}$$

We use the notation $S^{k,l} = S_k \cdot S_{k-1} \cdot \ldots \cdot S_l$ for $k \geq 0$ and $0 \leq l \leq k$. If $l > k$, then $S^{k,l} = E$ is identity operator. Then from (26) follows

$$\delta_{k+1} = S^{k,0}\delta_0 + \Delta\sum_{j=0}^{k} S^{k,j+1}\widehat{\delta}_j + \Delta\sum_{j=0}^{k} S^{k,j+1} d_j. \tag{28}$$

From (27), (28) and the definition of the stability of the method (23) follows

$$\|\delta_{k+1}\|_Y \le K\varDelta \sum_{j=0}^{k} \max_{j-m\le l\le j}\{\|\delta_l\|_Y\} + \|\delta_0\|_Y + (\theta - t_0) \max_{0\le l\le M-1}\{\|d_l\|_Y\}. \tag{29}$$

We denote

$$R_0 = \max_{-m\le l\le 0}\{\|\delta_l\|_Y\},\ R = \max_{0\le l\le M-1}\{\|d_l\|_Y\},\ D = R_0 + (\theta - t_0)R, \tag{30}$$

Then we transform estimate (29) as follows:

$$\|\delta_{k+1}\|_Y \le K\varDelta \sum_{j=0}^{n} \max_{j-m\le l\le j}\{\|\delta_l\|_Y\} + D. \tag{31}$$

Let us prove the following estimate by induction on $k = 1, \dots, M$.

$$\|\delta_k\|_Y \le D(1 + K\varDelta)^n. \tag{32}$$

Induction base. If we set $k = 0$ in (31), then

$$\|\delta_1\|_Y \le K\varDelta R_0 + D \le K\varDelta D + D = D(1 + K\varDelta).$$

Induction step. Suppose that the estimate (32) is true for all indices from 1 to k. We show that the estimate is also true for $k + 1$. Let us fix $j \le k$. Let $l_0 = l_0(j)$ be the index for which $\max\limits_{j-m\le l\le j}\{\|\delta_l\|_Y\}$ is achieved. There are two possible situations:

1. $l_0 \le 0$; then, $\max\limits_{j-m\le l\le j}\{\|\delta_l\|_Y\} = \|\delta_{l_0}\|_Y \le R_0 \le D(1 + K\varDelta)^j$.
2. $1 \le l_0 \le j$; then, by the induction assumption

$$\max_{j-m\le l\le j}\{\|\delta_l\|_Y\} = \|\delta_{l_0}\|_Y \le D(1 + K\varDelta)^{l_0} \le D(1 + K\varDelta)^j.$$

We get that in any case the following estimate is performed:

$$\max_{j-m\le l\le j}\{\|\delta_l\|_Y\} \le D(1 + K\varDelta)^j.$$

From the received estimate and (31) it follows

$$\begin{aligned}\|\delta_{k+1}\|_Y &\le K\varDelta \sum_{j=0}^{k} D(1 + K\varDelta)^j + D = D + K\varDelta D + K\varDelta \sum_{j=1}^{k} D(1 + K\varDelta)^j \\ &= D(1 + K\varDelta) + K\varDelta D(1 + K\varDelta)\frac{(1 + K\varDelta)^k - 1}{1 + K\varDelta - 1} \\ &= D(1 + K\varDelta) + D(1 + K\varDelta)((1 + K\varDelta)^k - 1) = D(1 + K\varDelta)^{k+1}.\end{aligned}$$

Therefore, the estimate (32) is proved, and from it we obtain the following inequality

$$\|\delta_k\|_Y \leq D\mathrm{e}^{K(\theta - t_0)}. \tag{33}$$

By definition (30) of the value D, the following inequality holds

$$D \leq C(\varDelta^{\min\{p_1, p_3\}} + h^{\min\{p_2, p_4\}}). \tag{34}$$

Then the conclusion of the theorem follows from (33), (34). □

5 The Embedding into the General Difference Scheme with Aftereffect

We embed scheme (6)–(8) into the general scheme. We define the values of the discrete model by the vector $y_k = (u_k^{000}, u_k^{001}, \ldots, u_k^{N_1N_2N_3})^T \in Y$ for every $t_k \in \Sigma_\varDelta$, where Y—vector space of dimension $q = (N_1 + 1)(N_2 + 1)(N_3 + 1)$, and T—the transposition symbol. We introduce in the space Y the operators $A_\beta(\bar{t}),\ A(\bar{t})$ using the notations from Sect. 3:

$$\overset{\circ}{\Lambda}_\beta\ (\bar{t})u_k^{i_1i_2i_3} = \begin{cases} 0,\ x^{i_1i_2i_3} \in \gamma_h, \\ \Lambda_\beta(\bar{t})u_k^{i_1i_2i_3},\ x^{i_1i_2i_3} \in \omega_h, \end{cases} \quad \beta = 1, 2, 3.$$

$$A_\beta(\bar{t})y_k = (-\overset{\circ}{\Lambda}_\beta\ (\bar{t})u_k^{000}, \ldots, -\overset{\circ}{\Lambda}_\beta\ (\bar{t})u_k^{N_1N_2N_3})^T,\ \beta = 1, 2, 3.$$
$$A(\bar{t}) = A_1(\bar{t}) + A_2(\bar{t}) + A_3(\bar{t}).$$

Using identities $y_{k+\frac{1}{3}} = y_k + \varDelta\dfrac{y_{k+\frac{1}{3}} - y_k}{\varDelta}$, $y_{k+\frac{2}{3}} = y_{k+\frac{1}{3}} + \varDelta\dfrac{y_{k+\frac{2}{3}} - y_{k+\frac{1}{3}}}{\varDelta}$, $y_{k+1} = y_{k+\frac{2}{3}} + \varDelta\dfrac{y_{k+1} - y_{k+\frac{2}{3}}}{\varDelta}$ and introducing the operators

$$B_\beta(\bar{t}) = E + \varDelta s A_\beta(\bar{t}),\ \beta = 1, 2, 3,$$

we bring the system (6)–(8) to the canonical form [8]:

$$B_1(\bar{t})\frac{y_{k+\frac{1}{3}} - y_k}{\varDelta} + A_1(\bar{t})y_k = F_k(v(\cdot)),\ k = 0, \ldots, M - 1, \tag{35}$$

$$B_2(\bar{t})\frac{y_{k+\frac{2}{3}} - y_{k+\frac{1}{3}}}{\varDelta} + A_2(\bar{t})y_k = F_k(v(\cdot)),\ k = 0, \ldots, M - 1, \tag{36}$$

$$B_3(\bar{t})\frac{y_{k+1}-y_{k+\frac{2}{3}}}{\Delta}+A_3(\bar{t})y_k=F_k(v(\cdot)).\ k=0,\dots,M-1, \text{ where} \tag{37}$$

$$F_k(v(\cdot))=(F_k^{000}\big(v_k^{000}(\cdot)\big),\dots,F_k^{N_1N_2N_3}\big(v_k^{N_1N_2N_3}(\cdot)\big))^T,\ v(\cdot)=I(\{y_l\}_k)\in Q^q[-\tau,\Delta].$$

Here, the interpolation space $V=Q^q[-\tau,\Delta]$ is the space of q-dimensional vector functions every component of which belongs to the space $Q[-\tau,\Delta]$. Since condition (4) holds, then for all $\bar{t}$ the operators $A(\bar{t})$, $A_\beta(\bar{t})$ is positive and self-adjoint [8] in the sense of the scalar product (12). The norm in the space Y is defined as $\|y\|_Y=\sqrt{(y,y)}$.

E (identity operator) is positive and self-adjoint in the sense of the scalar product (12). Thus, for all $\bar{t}$ the operators $B_\beta(\bar{t})$ are positive and self-adjoint in the sense of the scalar product (12). The operators $A_1(\bar{t})$ and $B_1(\bar{t})$, $A_2(\bar{t})$ and $B_2(\bar{t})$, $A_3(\bar{t})$ and $B_3(\bar{t})$ are commutative respectively for all $\bar{t}$ in the sense of the scalar product (12). Let us obtain the stability conditions. For this we use the fact that for all $\bar{t}$ following inequality holds $\|A_\beta(\bar{t})\|\le\dfrac{4c_2}{h_\beta^2}$. Let us consider

$$B_\beta(\bar{t})-\frac{\Delta}{2}A_\beta(\bar{t})=E+s\Delta A_\beta(\bar{t})-\frac{\Delta}{2}A_\beta(\bar{t})$$
$$=E+(s-\frac{1}{2})\Delta A_\beta(\bar{t})\ge 0 \text{ under } s\ge\frac{1}{2}-\frac{h_\beta^2}{4c_2\Delta},\ \beta=1,2,3. \tag{38}$$

We will consider $s\ge\dfrac{1}{2}-\dfrac{min(h_1^2,h_2^2,h_3^2)}{4c_2\Delta}$ to fulfill the condition (38) for all β. Since $B_\beta(\bar{t})$ is a positive operator in a finite-dimensional Hilbert space for all $\bar{t}$, then there exists $B_\beta^{-1}(\bar{t})$ for all $\bar{t}$ [8]. Therefore, we can transform the system (35)–(37) to explicit form

$$y_{k+\frac{1}{3}}=S_1(\bar{t})y_k+\Delta\Phi_1(\bar{t},I(\{y_l\}_k),\Delta),\ k=0,\dots,M-1, \tag{39}$$

$$y_{k+\frac{2}{3}}=S_2(\bar{t})y_{k+\frac{1}{3}}+\Delta\Phi_2(\bar{t},I(\{y_l\}_k),\Delta),\ k=0,\dots,M-1, \tag{40}$$

$$y_{k+1}=S_3(\bar{t})y_{k+\frac{2}{3}}+\Delta\Phi_3(\bar{t},I(\{y_l\}_k),\Delta),\ k=0,\dots,M-1, \tag{41}$$

where the transition operators $S_\beta(\bar{t})$ and the function of advance by a step are defined by the formulas

$$S_\beta(\bar{t})=E-\Delta B_\beta^{-1}(\bar{t})A_\beta(\bar{t}),\ \Phi_\beta(\bar{t},v,\Delta)=B_\beta^{-1}(\bar{t})F_k(v(\cdot)),\ \beta=1,2,3.$$

We exclude $y_{k+\frac{1}{3}}$ and $y_{k+\frac{2}{3}}$ from the system (39)–(41). Obtain

$$y_{k+1} = S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})y_k + \Delta[S_3(\bar{t})S_2(\bar{t})\Phi_1(\bar{t}, I(\{y_l\}_k, \Delta) + \\ +S_3(\bar{t})\Phi_2(\bar{t}, I(\{y_l\}_k), \Delta) + \Phi_3(\bar{t}, I(\{y_l\}_k), \Delta)], \ k = \overline{0, M-1}. \tag{42}$$

Let us investigate the stability of the resulting scheme. To this end, along with equation (42), consider the homogeneous difference scheme in the explicit form

$$y_{k+1} = S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})y_k, \ k = \overline{0, M-1}, \tag{43}$$

In view of positive and self-adjoint operators $A_\beta(\bar{t})$ and $B_\beta(\bar{t})$, and conditions (5), (38) we can apply stability criterion for two-layer difference schemes for the case of variable coefficient of heat conductivity [9]. We obtain

$$\|y_{k+\frac{1}{3}}\|_Y \le \|y_k\|_Y, \ \|y_{k+\frac{2}{3}}\|_Y \le \|y_{k+\frac{1}{3}}\|_Y, \ \|y_{k+1}\|_Y \le \|y_{k+\frac{2}{3}}\|_Y, \ k = \overline{0, M-1}. \tag{44}$$

Therefore, for the equivalent homogeneous equation (43) we have that

$$\|S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})y_k\|_Y \le \|y_k\|_Y, \ k = 0, \ldots, M-1. \tag{45}$$

We will consider $S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})$ as the operator S_k from Sect. 4. Then from (45) we obtain

$$\|S_k y_k\|_Y = \|S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})y_k\|_Y \le \|y_k\|_Y.$$

Consequently, we obtain for $k = 0, \ldots, M-1$

$$\|S_k\|_Y = \sup_{y_k \neq 0} \frac{\|S_k y_k\|_Y}{\|y_k\|_Y} \le 1, \text{ under condition } s > \frac{1}{2} - \frac{min(h_1^2, h_2^2, h_3^2)}{4c_2\Delta}. \tag{46}$$

Thus, under condition $s \ge \frac{1}{2} - \frac{min(h_1^2, h_2^2, h_3^2)}{4c_2\Delta}$ the scheme (42) is stable. Note that from (44) for $\beta = 1, 2, 3$ and all $\bar{t}$ we obtain the following inequality

$$\|S_\beta(\bar{t})\|_Y \le 1.$$

The function of exact values is defined as follows

$$z_k = (u(x^{000}, t_k), u(x^{001}, t_k), \ldots, u(x^{N_1N_2N_3}, t_k))^T \in Y, \ k = 0, \ldots, M.$$

We take the starting values of the model equal to the function of exact values:

$$y_j = z_j = (\varphi(x^{000}, t_j), \varphi(x^{001}, t_j), \ldots, \varphi(x^{N_1N_2N_3}, t_j))^T, \ j = -m, \ldots, 0.$$

The following theorem connects the definition of the residual without interpolation (17) in the method of fractional steps and the definition of the residual with interpolation (25) in the general scheme.

Theorem 3. *Suppose that the functions $F_k^{i_1i_2i_3}$ are Lipschitz, in addition, the interpolation-extrapolation operator I is Lipschitz and has order of error Δ^{p_0} on the exact solution. Suppose also that the residual in the sense of (17) has order $\Delta^{p_1} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4}$. Then, the residual with interpolation has order $\Delta^{\min\{p_1,p_0\}} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4}$.*

Proof. We write the definitions of the residual in the sense of (17) and (25)

$$\Psi_k^{i_1i_2i_3} = \frac{u(x^{i_1i_2i_3}, t_{k+1}) - S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})u(x^{i_1i_2i_3}, t_k)}{\Delta} - \frac{1}{3}[S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t})$$
$$+S_3(\bar{t})B_2^{-1}(\bar{t}) + B_3^{-1}(\bar{t})]f(x^{i_1i_2i_3}, \bar{t}, u(x^{i_1i_2i_3}, \bar{t}), u_{\bar{t}}(x^{i_1i_2i_3}, \cdot)).$$
$$d_k = \frac{z_{k+1} - S_3(\bar{t})S_2(\bar{t})S_1(\bar{t})z_k}{\Delta} - [S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t})$$
$$+S_3(\bar{t})B_2^{-1}(\bar{t}) + B_3^{-1}(\bar{t})]F(I(\{z_l\}_k)),\ k = 0, \dots, M-1.$$

For every component of the vector z we have

$$|d_k^{i_1i_2i_3}| = |\frac{u(x^{i_1i_2i_3}, t_{k+1}) - S_3^{i_1i_2i_3}(\bar{t})S_2^{i_1i_2i_3}(\bar{t})S_1^{i_1i_2i_3}(\bar{t})u(x^{i_1i_2i_3}, t_k)}{\Delta} - [B_3^{-1}(\bar{t})$$
$$+S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t}) + S_3(\bar{t})B_2^{-1}(\bar{t})]F^{i_1i_2i_3}(I(\{u(x^{i_1i_2i_3}, t_l)\}_k)),\ k = \overline{0, M-1}.$$

We add and subtract $f(x^{i_1i_2i_3}, \bar{t}, u(x^{i_1i_2i_3}, \bar{t}), u_{\bar{t}}(x^{i_1i_2i_3}, \cdot))$. Then

$$|d_k^{i_1i_2i_3}| = |\frac{u(x^{i_1i_2i_3}, t_{k+1}) - S_3^{i_1i_2i_3}(\bar{t})S_2^{i_1i_2i_3}(\bar{t})S_1^{i_1i_2i_3}(\bar{t})u(x^{i_1i_2i_3}, t_k)}{\Delta} - \frac{1}{3}[S_3(\bar{t})B_2^{-1}(\bar{t})$$
$$+S_3(\bar{t})S_2(\bar{t})B_1^{-1}(\bar{t}) + B_3^{-1}(\bar{t})][f(x^{i_1i_2i_3}, \bar{t}, u(x^{i_1i_2i_3}, \bar{t}), I(\{u(x^{i_1i_2i_3}, t_l)\}_{k+\frac{1}{2}}))$$
$$-f(x^{i_1i_2i_3}, \bar{t}, u(x^{i_1i_2i_3}, \bar{t}), u_{\bar{t}}(x^{i_1i_2i_3}, \cdot)) + f(x^{i_1i_2i_3}, \bar{t}, u(x^{i_1i_2i_3}, \bar{t}), u_{\bar{t}}(x^{i_1i_2i_3}, \cdot))].$$

It is easy to prove that (24) implies the following componentwise inequalities

$$|S_\beta^{i_1i_2i_3}(\bar{t})y_k^{i_1i_2i_3}| \le \|S\|_Y|y_k^{i_1i_2i_3}|,\ \beta = 1, 2, 3,\ y_k = (y_k^{000}, \dots, y_k^{N1N2N3}). \quad (47)$$

We obtain following inequalities from conditions (16), (46) and (47), and the facts that F is Lipschitz with respect to $u(\cdot)$, and the definition of the order of the residual in the sense of (17), and the definition of the order of an interpolation-extrapolation operator

$$|d_k^{i_1i_2i_3}| \le C_\Psi(\Delta^{p_1} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4}) + L_f C_2 \Delta^{p_0}$$
$$\le (C_\Psi + L_f C_2)(\Delta^{\min\{p_1,p_0\}} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4}),\ k = \overline{0, M-1}.$$

Therefore, the norm of the vector d_k satisfies the inequality

$$\|d_k\|_Y = \sqrt{(d_k, d_k)} = (\sum_{i_1=0}^{N_1}\sum_{i_2=0}^{N_2}\sum_{i_3=0}^{N_3}(d_k^{i_1 i_2 i_3})^2 h_1 h_2 h_3)^{\frac{1}{2}}$$

$$\le (C_\Psi + L_f C_2)(\Delta^{\min\{p_1,p_0\}} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4})(\sum_{i_1=0}^{N_1}\sum_{i_2=0}^{N_2}\sum_{i_3=0}^{N_3} h_1 h_2 h_3)^{\frac{1}{2}}$$

$$\le \sqrt{2X_1X_2X_3}(C_\Psi + L_f C_2)(\Delta^{\min\{p_1,p_0\}} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4}).$$

The conclusion of the theorem follows from this relation. □

Thus, the embedding of the method of fractional steps for the multidimensional heat conduction equation into the general scheme is complete. Based on theorem 2 we obtain the following theorem.

Theorem 4. *Suppose that condition* $s \ge \frac{1}{2} - \frac{min(h_1^2, h_2^2, h_3^2)}{4c_2\Delta}$ *holds. Suppose also that the residual in the sense of (17) has order* $\Delta^{p_1} + h_1^{p_2} + h_2^{p_3} + h_3^{p_4}$*. Also, the functions* F_k^i *are Lipschitz, and the interpolation-extrapolation operator I is Lipschitz and has order of error* Δ^{p_0} *on the exact solution. Then, the method converges with order* $\Delta^{\min\{p_1,p_0\}+h_1^{p_2}+h_2^{p_3}+h_3^{p_4}}$.

Using this theorem, we conclude that the method of fractional steps with a piecewise constant interpolation and extrapolation by expansion has order $\Delta + h_1^2 + h_2^2 + h_3^2$.

6 Numerical Example

We will solve test example by means the method of fractional steps for the parameter $s = 1/2$. To more clearly demonstrate the operation of the method, we will consider two-dimensional heat conduction equation with aftereffect. However, one can consider examples of equations with a large number of spatial variables. Let us consider the following test equation with variable concentrated and distributed delays:

$$\frac{\partial u(x,y,t)}{\partial t} = \frac{\partial}{\partial x}(K_1(x,y,t)\frac{\partial u(x,y,t)}{\partial x}) + \frac{\partial}{\partial y}(K_2(x,y,t)\frac{\partial u(x,y,t)}{\partial y})$$

$$+(y - xt^2)(2 + sin(x)) - t^2y^2(cos(x) - xsin(x)) - \sqrt{\frac{t}{2}(2 + sin(x))y}$$

$$+\sqrt{u(x,y,t-\tau(t))} + \int_{-\tau(t)}^{0}(x+y)u(x,y,t+s)ds - \frac{3}{8}(x+y)(2+sin(x))yt^2, \quad (48)$$

Table 1 The maximum in time of the norm of difference between the approximate and exact solutions

	M=20	M=200	M=2000
$N_1 = 10, N_2 = 10$	0.2913	0.0357	0.0106
$N_1 = 20, N_2 = 20$	0.3068	0.0325	0.0049
$N_1 = 40, N_2 = 40$	0.3155	0.0323	0.0036

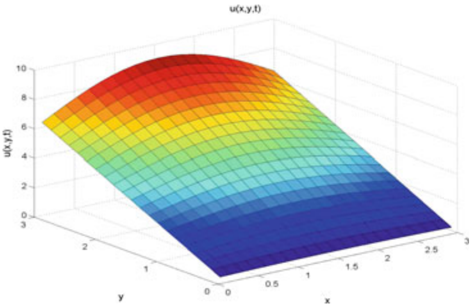


Fig. 1 The approximate solution for $t = 1$. The number of grid points in t: $M = 200$, in x: $N_1 = 20$, in y: $N_2 = 20$

where $x \in [0.1, 3]$, $y \in [0.1, 3]$, $t \in [0.1, 1]$, $\tau(t) = t/2$. Coefficients of heat conductivity are taken as follows: $K_1(x, y, t) = K_2(x, y, t) = xyt$. The following initial conditions are given:

$$u(x, y, t) = (2 + sin(x))yt, \;\; x \in [0.1, 3], \;\; y \in [0.1, 3], \;\; t \in [0.05, 0.1].$$

The following boundary conditions are also given:

$$\begin{aligned} u(0.1, y, t) &= (2 + sin(0.1))yt, \;\; y \in [0.1, 3], \;\; t \in [0.1, 1], \\ u(3, y, t) &= (2 + sin(3))yt, \;\; y \in [0.1, 3], \;\; t \in [0.1, 1], \\ u(x, 0.1, t) &= 0.1(2 + sin(x))t, \;\; x \in [0.1, 3], \;\; t \in [0.1, 1], \\ u(x, 3, t) &= 3(2 + sin(x))t, \;\; x \in [0.1, 3], \;\; t \in [0.1, 1]. \end{aligned}$$

The exact solution is the function $u(x, y, t) = (2 + sin(x))yt$. The following are the results of the numerical experiment. The Table 1 contains the comparison of the maximum in time of the norm of difference between the approximate and exact solutions of the equation (48) for different steps h_1, h_2, Δ. The approximate solution of the equation (48) is shown in Figure 1 for $t = 1$.

Acknowledgements This work was supported by RFBR Grant 19-01-00019 and Act 211 Government of the Russian Federation, contract 02.A03.21.0006.

References

1. Castro, M.A., Rodriguez, F., Cabrera, J., Martin, J.A.: Difference schemes for time-dependent heat conduction models with delay. Int. J. Comput. Math. **91**(1), 53–61 (2014)
2. Garcia, P., Castro, M.A., Martin, J.A., Sirvent, A.: Numerical solutions of diffusion mathematical models with delay. Math. Comput. Model. **50**(5–6), 860–868 (2013)
3. Kropielnicka, K.: Convergence of implicit difference methods for parabolic functional differential equations. Int. J. Mat. Anal. **1**(6), 257–277 (2007)
4. Lekomtsev, A.V., Pimenov, V.G.: Convergence of the alternating direction methods for the numerical solution of a heat conduction equation with delay. Proc. Steklov Inst. Math. **272**(1), 101–118 (2011)
5. Lekomtsev, A.V., Pimenov, V.G.: Convergence of the scheme with weights for the numericalsolution of a heat conduction equation with delay for the case of variable coefficient of heatconductivity. Appl. Math. Comput. **256**, 83–93 (2015)
6. Pimenov, V.G.: General linear methods for numerical solving functional-differential equations. Differ. Equ. **37**(1), 116–127 (2001)
7. Pimenov, V.G., Lozhnikov, A.B.: Difference schemes for the numerical solution of the heat conduction equation with aftereffect. Proc. Steklov Inst. Math. **275**(S1), 137–148 (2011)
8. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker, New York (2001)
9. Samarskii, A.A., Gulin, A.V.: Stability of Difference Schemes. URSS, Moscow (2009). [in Russian]
10. Skeel, R.D.: Analysis of fixed-stepsize methods. SIAM J. Numer. Anal. **13**(5), 664–685 (1976)
11. Tavernini, L.: Finite difference approximations for a class of semilinear volterra evolution problems. SIAM J. Numer. Anal. **14**(5), 931–949 (1977)
12. Van der Houwen, P.J., Sommeijer, B.P., Baker, C.T.H.: On the stability of predictor-corrector methods for parabolic equations with delay. IMA J. Numer. Anal. **6**(1), 1–23 (1986)
13. Wu, J.: Theory and Applications of Partial Functional Differential Equations. Springer, New York (1996)
14. Zhang, B., Zhou, Y.: Qualitative Analysis of Delay Partial Difference Equations. Hindawi Publishing Corporation, New York (2007)
15. Zubik-Kowal, B.: The method of lines for parabolic differential-functional equations. IMA J. Numer. Anal. **17**(1), 103–123 (1997)

Hyers-Ulam Stability of a Nonlinear Volterra Integral Equation on Time Scales

Andrejs Reinfelds and Shraddha Christian

Abstract We study Hyers-Ulam stability of a nonlinear Volterra integral equation on unbounded time scales. Sufficient conditions are obtained based on the Banach fixed point theorem and Bielecki type norm.

Keywords Hyers-Ulam stability · Nonlinear Volterra integral equation · Unbounded time scales

1 Introduction

In 1940 S.M. Ulam [23] at the University of Wisconsin raised the question when a solution of an equation, differing slightly from a given one, must be somehow near to the exact solution of the given equation. In the following year, D.H. Hyers [10] gave an affirmative answer to the question of S.M. Ulam for additive Cauchy equation in a Banach space. So the stability concept proposed by S.M. Ulam and D.H. Hyers, was named as *Hyers-Ulam stability*. Afterwards Th.M. Rassias [15] introduced new ideas of Hyers-Ulam stability using unbounded right-hand sides in the involved inequalities, depending on certain functions, introducing therefore the so-called *Hyers-Ulam-Rassias stability*.

In 2007, S.M. Jung [13] proved, using a fixed point approach, that the Volterra nonlinear integral equation is Hyers-Ulam-Russias stable, on a compact interval under certain conditions. Then several authors [5, 11, 12] generalized the previous result on the Volterra integral equations to infinite interval in the case when the

A. Reinfelds (✉)
Institute of Mathematics and Computer Science, University of Latvia,
29 Raiņa bulvāris, Rīga 1459, Latvia
e-mail: reinf@latnet.lv

A. Reinfelds · S. Christian
Department of Mathematics, University of Latvia,
3 Jelgavas iela, Rīga 1004, Latvia
e-mail: sc16024@lu.lv

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_10

integrand is Lipschitz with a fixed Lipschitz constant. In the near past many research papers have been published about Ulam-Hyers stability of Voltera integral equations of different type including nonlinear Volterra integro-differential equations, mixed integral dynamic system with impulses etc. [6, 7, 18, 21].

The theory of time scales analysis has been rising fast and has acknowledged a lot of interest. The pioneer of this theory was S. Hilger [8]. He introduced this theory in 1988 with the inspiration to unify continuous and discrete calculus. For the introduction to the calculus on time scales and to the theory of dynamic equations on time scales, we recommend the books [3] and [4] by M. Bohner and A. Peterson.

T. Kulik and C.C. Tisdell [14, 22] gave the basic qualitative and quantitative results to Volterra integral equations on time scales in the case when the integrand is Lipschitz with a fixed Lipschitz constant. A. Reinfelds and S. Christian [16, 17] generalized previous results using Lipschitz functions, whose Lipschitz coefficients can be unbounded.

To the best of our knowledge, the first ones who pay attention to Hyers-Ulam stability for Volterra integral equations on time scales are S. Andras, A.R. Meszaros [1] and L. Hua, Y. Li, J. Feng [9]. However they restricted their research to the case when integrand satisfies Lipschitz conditions with some Lipschitz constant. We generalize the results of [1, 9] using Lipschitz functions, whose Lipschitz coefficients can be an unbounded, and the Banach's fixed point theorem at appropriate functional space with Bielecki type norm. There are also papers on impulsive integral equations on time scales [19, 20].

2 Notations and Preliminaries

A time scale $\mathbb{T}$ is an arbitrary non empty closed subset of the real numbers $\mathbb{R}$. Since a time scale may or may not be connected, the concept of jump operator is useful for describing the structure of the time scale under consideration and is also used in defining the delta derivative. The forward jump operator $\sigma : \mathbb{T} \to \mathbb{T}$ is defined by the equality

$$\sigma(t) = \inf\{s \in \mathbb{T} \mid s > t\}$$

while the backward jump operator $\rho : \mathbb{T} \to \mathbb{T}$ is defined by the equality

$$\rho(t) = \sup\{s \in \mathbb{T} \mid s < t\}.$$

We define the graininess function $\mu : \mathbb{T} \to [0, +\infty)$ by the relation

$$\mu(t) = \sigma(t) - t.$$

The jump operators allow the classification of points in a time scale $\mathbb{T}$. If $\sigma(t) > t$, then the point $t \in \mathbb{T}$ is called right scattered while if $\rho(t) < t$, then the point $t \in \mathbb{T}$ is called left scattered. If $\sigma(t) = t$ then $t \in \mathbb{T}$ is called right dense while if $\rho(t) = t$ then $t \in \mathbb{T}$ is called left dense.

The function $g\colon \mathbb{T} \to \mathbb{R}$ is called rd-continuous provided it is continuous at every right dense points in $\mathbb{T}$ and its left sided limits exist at every left dense points in $\mathbb{T}$. The function $g\colon \mathbb{T} \to \mathbb{R}$ is regressive if

$$1 + \mu(t)g(t) \neq 0 \quad \text{for all} \quad t \in \mathbb{T}.$$

Assume $g\colon \mathbb{T} \to \mathbb{R}$ is a function and fix $t \in \mathbb{T}^{\kappa}$. The delta derivative (also Hilger derivative) $g^{\Delta}(t)$ exists if for every $\varepsilon > 0$ there exists a neighbourhood $U = (t - \delta, t + \delta) \bigcap \mathbb{T}$ for some $\delta > 0$ such that

$$\left|(g(\sigma(t)) - g(s)) - g^{\Delta}(t)(\sigma(t) - s)\right| \leq \varepsilon \left|\sigma(t) - s\right|, \text{ for all } s \in U.$$

Take $\mathbb{T} = \mathbb{R}$ and g is differentiable in the ordinary sense at $t \in \mathbb{T}$. Then $g^{\Delta}(t) = g'(t)$ is the derivative used in standard calculus. Take $\mathbb{T} = \mathbb{Z}$. Then $g^{\Delta}(t) = \Delta g(t)$ is the forward difference operator used in difference equation.

If $F^{\Delta}(t) = g(t)$ then define the (Cauchy) delta integral by

$$\int_r^s g(t)\, \Delta t = F(s) - F(r), \text{ for all } r, s \in \mathbb{T}.$$

If $\mathbb{T} = \mathbb{R}$, then

$$\int_r^s g(t)\, \Delta t = \int_r^s g(t)\, \mathrm{d}t$$

while $\mathbb{T} = \mathbb{Z}$, then

$$\int_r^s g(t)\, \Delta t = \sum_{t=r}^{s-1} g(t), \text{ if } r, s \in \mathbb{Z} \text{ and } r < s.$$

Let $\beta\colon \mathbb{T} \to \mathbb{R}$ be a nonnegative (and therefore regressive) and rd-continuous scalar function. The Cauchy initial value problem for scalar linear equation

$$x^{\Delta} = \beta(t)x, \quad x(a) = 1, \quad a \in \mathbb{T}$$

has the unique solution $\mathrm{e}_{\beta}(\cdot, a)\colon \mathbb{T} \to \mathbb{R}$ [3]. More explicitly, using the cylinder transformation the exponential function $\mathrm{e}_{\beta}(\cdot, a)$ is given by

$$\mathrm{e}_{\beta}(t, a) = \exp\left(\int_a^t \xi_{\mu(s)}(\beta(s))\, \Delta s\right),$$

where

$$\xi_h(z) = \begin{cases} z, & h = 0 \\ \frac{1}{h}\log(1 + hz), & h > 0. \end{cases}$$

Observe that we also have Bernoulli's type estimate [2]

$$1+\int_a^t \beta(s)\,\Delta s \le \mathrm{e}_\beta(t,a) \le \exp\left(\int_a^t \beta(s)\,\Delta s\right) \tag{1}$$

for all $t \in I_{\mathbb{T}} = [a, +\infty) \cap \mathbb{T}$.

Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^n$. We will consider the linear space of continuous functions $C(I_{\mathbb{T}}; \mathbb{R}^n)$ such that,

$$\sup_{t\in I_{\mathbb{T}}} \frac{|x(t)|}{\mathrm{e}_\beta(t,a)} < \infty$$

and denote this special space by $C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$. The space $C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$ endowed with Bielecki type norm

$$\|x\|_\beta = \sup_{t\in I_{\mathbb{T}}} \frac{|x(t)|}{\mathrm{e}_\beta(t,a)}$$

is a Banach space.

3 Hyers-Ulam Stability of Nonlinear Volterra Integral Equation on Unbounded Time Scales

Consider the nonlinear Volterra integral equation

$$x(t) = f(t) + \int_a^t K(t,s,x(s))\,\Delta s, \quad a, t \in I_{\mathbb{T}} = [a, +\infty) \cap \mathbb{T}. \tag{2}$$

In paper [16], we proved the existence and uniqueness of solution of (2) using Lipschitz functions, whose Lipschitz coefficients can be unbounded.

Theorem 1. *Let $K\colon I_{\mathbb{T}} \times I_{\mathbb{T}} \times \mathbb{R}^n \to \mathbb{R}^n$ be jointly continuous in its first and third variables and rd-continuous in its second variable, $f\colon I_{\mathbb{T}} \to \mathbb{R}^n$ be continuous, $L\colon I_{\mathbb{T}} \to \mathbb{R}$ be rd-continuous, $\gamma > 1$ and $\beta(s) = L(s)\gamma$. If*

$$|K(t,s,x) - K(t,s,x')| \le L(s)|x - x'|, \quad x, x' \in \mathbb{R}^n, \quad s < t, \tag{3}$$

$$m = \sup_{t\in I_{\mathbb{T}}} \frac{1}{\mathrm{e}_\beta(t,a)} \left| f(t) + \int_a^t K(t,s,0)\,\Delta s \right| < \infty,$$

then the nolinear Volterra integral Eq. (2) has a unique solution $x \in C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$.

Consider the Banach space $C_\beta(I_\mathbb{T}; \mathbb{R}^n)$. To prove the Theorem 1 we define an operator $F\colon C_\beta(I_\mathbb{T}; \mathbb{R}^n) \to C_\beta(I_\mathbb{T}; \mathbb{R}^n)$ by expression

$$[Fx](t) = \int_a^t (K(t, s, x(s)) - K(t, s, 0))\, \Delta s.$$

Here $L\colon I_\mathbb{T} \to \mathbb{R}$ is the Lipschitz type function defined by (3) and $\beta(s) = L(s)\gamma$, where $\gamma > 1$. Analogously to the Theorem 1 [16] we can verify that for any $x, x' \in C_\beta(I_\mathbb{T}; \mathbb{R}^n)$

$$\begin{aligned}
\|[Fx](t) - [Fx'](t)\|_\beta &= \sup_{t\in I_\mathbb{T}} \frac{|[Fx](t) - [Fx'](t)|}{\mathrm{e}_\beta(t, a)} \\
&\le \sup_{t\in I_\mathbb{T}} \frac{1}{\mathrm{e}_\beta(t, a)} \int_a^t |K(t, s, x(s)) - K(t, s, x'(s))|\, \Delta s \\
&\le \sup_{t\in I_\mathbb{T}} \frac{1}{\mathrm{e}_\beta(t, a)} \int_a^t L(s)|x(s) - x'(s)|\, \Delta s \\
&= \sup_{t\in I_\mathbb{T}} \frac{1}{\mathrm{e}_\beta(t, a)} \int_a^t L(s)\mathrm{e}_\beta(s, a) \frac{|x(s) - x'(s)|}{\mathrm{e}_\beta(s, a)}\, \Delta s \\
&\le \|x - x'\|_\beta \sup_{t\in I_\mathbb{T}} \frac{1}{\mathrm{e}_\beta(t, a)} \int_a^t L(s)\mathrm{e}_\beta(s, a)\, \Delta s \\
&= \frac{\|x - x'\|_\beta}{\gamma} \sup_{t\in I_\mathbb{T}} \frac{1}{\mathrm{e}_\beta(t, a)} \int_a^t \gamma L(s)\mathrm{e}_\beta(s, a)\, \Delta s \\
&= \frac{\|x - x'\|_\beta}{\gamma} \sup_{t\in I_\mathbb{T}} \frac{1}{\mathrm{e}_\beta(t, a)} \int_a^t \mathrm{e}_\beta^\Delta(s, a)\, \Delta s \\
&= \frac{\|x - x'\|_\beta}{\gamma} \sup_{t\in I_\mathbb{T}} \left[1 - \frac{1}{\mathrm{e}_\beta(t, a)}\right] \\
&\le \frac{\|x - x'\|_\beta}{\gamma}.
\end{aligned}$$

So we get

$$\left\| \int_a^t K(t, s, x(s))\, \Delta s - \int_a^t K(t, s, x'(s))\, \Delta s \right\|_\beta \le \frac{\|x(t) - x'(t)\|_\beta}{\gamma}.$$

Definition 1. We say that integral Eq. (2) is Hyers-Ulam stable if there exists a constant $C > 0$ such that for each real number $\varepsilon > 0$ and for each solution $x \in C_\beta(I_\mathbb{T}; \mathbb{R}^n)$ of the inequality

$$\|x(t) - f(t) - \int_a^t K(t, s, x(s))\, \Delta s\|_\beta \le \varepsilon,$$

there exists a solution $x_0 \in C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$ of the integral Eq. (2) with the property

$$\|x(t) - x_0(t)\|_\beta \leq C\varepsilon.$$

□

Let us find sufficient conditions for the Hyers-Ulam stability of nonlinear Volterra integral equation on time scales.

Theorem 2. *Consider the nonlinear Volterra integral Eq. (2) satisfying conditions of Theorem 1. Suppose $x \in C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$ is such that satisfies the inequality*

$$\|x(t) - f(t) - \int_a^t K(t, s, x(s))\, \Delta s\|_\beta \leq \varepsilon.$$

Then nonlinear Volterra integral Eq. (2) is Hyers-Ulam stable.

Proof. According to the Theorem 1 [16], there is a unique solution x_0 of the Volterra integral Eq. (2) in Banach space $x_0 \in C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$. Therefore we get the estimate

$$\|x(t) - x_0(t)\|_\beta \leq \left\| x(t) - f(t) - \int_a^t K(t, s, x(s))\, \Delta s \right\|_\beta$$

$$+ \left\| \int_a^t K(t, s, x(s))\, \Delta s - \int_a^t K(t, s, x_0(s))\, \Delta s \right\|_\beta \leq \varepsilon + \gamma^{-1}\|x(t) - x_0(t)\|_\beta.$$

Hence,

$$\|x(t) - x_0(t)\|_\beta \leq C\varepsilon, \tag{4}$$

where $C = (1 - \gamma^{-1})^{-1}$. □

Example 1. Consider the scalar Volterra integral equation for an arbitrary $\mathbb{T}$

$$x(t) = t^2 + \int_a^t (s + \sigma(s))[x(s)^2 + 1]^{\frac{1}{2}}\, \Delta s, \quad a, t \in I_{\mathbb{T}} = [a, +\infty) \cap \mathbb{T}, a \geq 0. \tag{5}$$

According to Theorem 1 [16], there is a unique solution of the Volterra integral Eq. (5) in Banach space $C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$, where $\beta(t) = L(s)\gamma$ and $\gamma > 1$. Let us note that according to [2] we have estimate

$$1 + \gamma(t^2 - a^2) \leq \mathrm{e}_\beta(t, a) \leq \exp(\gamma(t^2 - a^2))$$

which ensures the existence of a solution.

It follows from the Theorem 2 that integral Eq. (5) is Hyers-Ulam stable in Banach space $C_\beta(I_{\mathbb{T}}; \mathbb{R}^n)$.

4 Hyers-Ulam Stability of Nonlinear Volterra Integral Equation on Bounded Time Scales

In the case of a bounded (compact) time scales $a, b \in I_{\mathbb{T}} = [a, b] \cap \mathbb{T}$ we have

$$1 \leq \sup_{t \in I_{\mathbb{T}}} \mathrm{e}_{\beta}(t, a) \leq \sup_{t \in I_{\mathbb{T}}} \exp \int_a^t \beta(s)\, \Delta s = M < \infty.$$

Let us note that every rd-continuous function on a compact interval is bounded. Therefore supremum norm and Bielecki type norm at Banach space $C_{\beta}(I_{\mathbb{T}}; \mathbb{R}^n)$ are equivalent

$$\sup_{t \in I_{\mathbb{T}}} |x(t)| \leq M \|x\|_{\beta} \leq M \sup_{t \in I_{\mathbb{T}}} |x(t)|.$$

We can take also $\gamma = 1$. Then $\beta(t) = L(t)$ and we get estimate

$$\|[Fx](t) - [Fx'](t)\|_{\beta} \leq (1 - M^{-1})\|x - x'\|_{\beta}.$$

From Theorem 2, we get

$$\|x(t) - x_0(t)\|_{\beta} \leq M \left\| x(t) - f(t) - \int_a^t K(t, s, x(s))\, \Delta s \right\|_{\beta}.$$

It follows

$$\begin{aligned} \sup_{t \in I_{\mathbb{T}}} |x(t) - x_0(t)| &\leq M \|x(t) - x_0(t)\|_{\beta} \\ &\leq M^2 \left\| x(t) - f(t) - \int_a^t K(t, s, x(s))\, \Delta s \right\|_{\beta} \\ &\leq M^2 \sup_{t \in I_{\mathbb{T}}} \left| x(t) - f(t) - \int_a^t K(t, s, x(s))\, \Delta s \right|. \end{aligned}$$

Here $C = M^2$.

It follows that integral Eq. (2) on bounded time scales is also Hyers-Ulam stable in Banach space with supremum norm.

5 Remarks

It might be interesting to obtain general results of Hyers-Ulam stability for functional, integral functional, operatorial equations etc., namely for

$$x(t) = F(t, x(t)), \quad t \in \mathbb{T}$$

in a Banach space endowed with appropriate Bielecki type norm.

References

1. Andras, S., Meszaros, A.R.: Ulam-Hyers stability of dynamic equations on time scales via Picard operators. Appl. Math. Comput. **219**(9), 4853–4864 (2013)
2. Bohner, M.: Some oscillation criteria for first order delay dynamic equations. Far East J. Appl. Math. **18**(3), 289–304 (2005)
3. Bohner, M., Peterson, A.: Dynamic Equations on Time Scales. An Introduction with Applications. Birkhäuser, Boston, Basel, Berlin (2001)
4. Bohner, M., Peterson, A.: Advances in Dynamic Equations on Time Scales. Birkhäuser, Boston, Basel, Berlin (2003)
5. Castro, L.P., Ramos, A.: Hyers-Ulam-Russias stability for a class of nonlinear Volterra integral equations. Banach J. Math. Anal. **3**(1), 36–43 (2009)
6. Castro, L.P., Simoes, A.M.: Different types of Hyers-Ulam-Rassias stabilities for a class of integro-differential equations. Filomat **31**(17), 5379–5390 (2017)
7. Castro, L.P., Simoes, A.M.: Hyers-Ulam-Rassias stability of nonlinear integral equations through the Bielecki metric. Math. Meth. Appl. Sci. **41**(17), 7367–7383 (2018)
8. Hilger, S.: Analysis on measure chains. A unified approach to continuous and discrete calculus. Results Math. **18**(1–2), 18–56 (1990)
9. Hua, L., Li, Y., Feng, J.: On Hyers-Ulam stability of dynamic integral equation on time scales. Math. Aeterna **4**(6), 559–571 (2014)
10. Hyers, D.H.: On the stability of linear functional equation. Proc. Nat. Acad. Sci. U.S.A. **27**(4), 222–224 (1941)
11. Gachpazan, M., Baghani, O.: Hyers-Ulam stability of Volterra integral equations. Int. J. Nonlinear Anal. Appl. **1**(2), 19–25 (2010)
12. Gavruta, P., Gavruta, L.: A new method for the generalized Hyers-Ulam-Rassias stability. Int. J. Nonlinear Anal. Appl. **1**(2), 11–18 (2010)
13. Jung, S.M.: A fixed point approach to the stability of a Volterra integral equation. Fixed Point Theory Appl. **2007** (2007). Article ID 57064
14. Kulik, T., Tisdell, C.C.: Volterra integral equations on time scales. Basic qualitative and quantitative results with applications to initial value problems on unbounded domains. Int. J. Difference Equ. **3**(1), 103–133 (2008)
15. Rassias, Th.M.: On the stability of the linear mapping in Banach spaces. Proc. Amer. Math. Soc. **72**(2), 297–300 (1978)
16. Reinfelds, A., Christian, S.: Volterra integral equations on unbounded time scales. Int. J. Difference Equ. **14**(2), 169–177 (2019)
17. Reinfelds, A., Christian, S.: A nonstandard Volterra integral equation on time scales. Demonstr. Math. **52**(1), 503–510 (2019)
18. Sevgin, S., Sevli, H.: Stability of a nonlinear Volterra integro-differential equation via a fixed point approach. Nonlinear Sci. Appl. **9**(1), 200–207 (2016)

19. Shah, S.O., Zada, A.: Existence, uniqueness and stability of solution to mixed integral dynamic systems with instantaneous and noninstantaneous impulses on time scales. Appl. Math. Comput. **359**, 202–213 (2019)
20. Shah, S.O., Zada, A., Hamza, A.E.: Stability analysis of the first order non-linear impulsive time varying delay dynamic system on time scales. Qual. Theory Dyn. Syst. **18**(3), 825–840 (2019)
21. Zada, A., Riaz, U., Khan, F.U.: Hyers-Ulam stability of impulsive integral equations. Bolletino dell'Unione Math. Ital. **12**(3), 453–467 (2019)
22. Tisdeil, C.C., Zaidi, A.: Basic qualitative and quantitative results for solutions to nonlinear dynamic equations on time scales with an application to economic modelling. Nonlinear Anal. **68**, 3504–3524 (2008)
23. Ulam, S.M.: A Collection of the Mathematical Problems. Interscience, New York (1960)

On Some Stochastic Algorithms for the Numerical Solution of the First Boundary Value Problem for the Heat Equation

Alexander S. Sipin and Andrey N. Kuznetsov

Abstract We deal with statistical modeling algorithms for the numerical solution of the first boundary value problem for the heat equation. Unbiased estimators of the solution of a boundary value problem are built on the trajectories of random walks. We consider a random walk on the boundary and a random walk on the cylinders inside the region in which the boundary problem must be solved. The results of computational experiments and some applications are presented. The complexity of algorithms are estimated numerically.

1 Introduction

Let D be a bounded domain in $\mathbf{R}^n$. We suppose that its boundary Γ consists of the any finite smooth parts.

Let $u(t, x)$ be a classical solution of the BVP for the heat equation

$$\begin{aligned} \frac{\partial u(t,x)}{\partial t} &= \Delta u(t,x) + f(t,x), \quad x \in D, \quad t > 0 \\ u(t,x) &= \psi(t,x), \quad x \in \Gamma, \quad t > 0 \\ u(0,x) &= \varphi(x), \quad x \in D, \end{aligned} \tag{1}$$

where $\varphi \in C(\overline{D})$, $\psi \in C([0,\infty) \times \Gamma)$, $f \in L_2([0,\infty) \times D)$ and $\varphi(x) = \psi(0,x)$ for $x \in \Gamma$.

In the numerical solution of boundary value problems, various methods are used, including probabilistic ones. Any probabilistic algorithm is based on a representation of the solution of a boundary value problem in the form of a mathematical expectation of some random variable. This random variable is called an unbiased estimator for

A. S. Sipin (✉) · A. N. Kuznetsov
Vologda State University, Lenina, 15, Vologda, Russia
e-mail: cac1909@mail.ru

A. N. Kuznetsov
e-mail: pmqqkan@mail.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_11

the $u(t, x)$. Unbiased estimators for the $u(t, x)$ are usually constructed on trajectories of random walks or random processes with continuous time. Examples of unbiased estimators for various boundary value problems can be found in books [1–4] and papers [5–7].

We use the Monte Carlo method to calculate the $u(t, x)$ and construct the unbiased estimators for the $u(t, x)$ on the trajectories of two Markov chains: the Random Walk on the cylinders (RWC) and Random Walk on the boundary (RWB). The complexity of algorithms is estimated numerically.

2 Random Walk on the Cylinders

Let us give a brief description of the walk simulation procedure and the construction of an unbiased estimator (for a detailed description, see [3]).

We use the Mean Value Theorem to construct RWC and estimator for the $u(t, x)$. Let $R = R(x)$ be the distance from a point $x \in D$ to the boundary Γ, D_R be the ball centered at x of radius R and let $r = \|y - x\|$. If $R^2 \geq 2nt$, then using Green formula for the function $u(\tau, y)$ and the function

$$v(\tau, y) = \frac{1}{[4\pi(t-\tau)]^{\frac{n}{2}}} \cdot \left(exp\left(-\frac{r^2}{4(t-\tau)}\right) - exp\left(-\frac{R^2}{4(t-\tau)}\right)\right)$$

we have Mean Value Theorem for the solution $u(t, x)$ of the problem (1)

$$u(t, x) = \int_0^t \int_{D_R} v(\tau, y) f(\tau, y) dy d\tau$$

$$+ \int_0^t \int_{D_R} \left(\frac{R^2}{4(t-\tau)^2} - \frac{n}{2(t-\tau)}\right) \frac{1}{[4\pi(t-\tau)]^{\frac{n}{2}}} \cdot exp\left(-\frac{R^2}{4(t-\tau)}\right) u(\tau, y) dy d\tau$$

$$+ \int_{D_R} v(0, y)\varphi(y) dy$$

$$+ \int_0^t \int_{\partial D_R} \frac{R}{2(t-\tau)} \frac{1}{[4\pi(t-\tau)]^{\frac{n}{2}}} \cdot exp\left(-\frac{R^2}{4(t-\tau)}\right) u(\tau, y) d_y S d\tau,$$

that can be written in a probability form:

$$u(t, x) = I_1 + I_2 + I_3 + I_4, \tag{2}$$

where

$$I_1 = E\mathbf{1}_{\left\{\gamma \leq \frac{R^2}{4t\theta}\right\}} t \cdot f\left(t - t\theta, x + 2\rho\sqrt{t\theta\gamma}\Omega\right), \tag{3}$$

$$I_2 = E\mathbf{1}_{\left\{\gamma > \frac{R^2}{4t}\right\}}\mathbf{1}_{\left\{\theta > \frac{n}{2\gamma}\right\}}u\left(t - \frac{R^2}{4\gamma}, x + R\rho\Omega\right), \tag{4}$$

$$I_3 = E\mathbf{1}_{\left\{\gamma \le \frac{R^2}{4t}\right\}}\varphi(x + 2\rho\sqrt{t\gamma}\Omega), \tag{5}$$

$$I_4 = E\mathbf{1}_{\left\{\gamma > \frac{R^2}{4t}\right\}}\mathbf{1}_{\left\{\theta \le \frac{n}{2\gamma}\right\}}u\left(t - \frac{R^2}{4\gamma}, x + R\Omega\right). \tag{6}$$

Here, $\mathbf{1}_A$ is the indicator function of the event A. Random vector Ω is uniformly distributed on the sphere of unit radius centered at the origin. The random variable θ is uniformly distributed on the segment [0, 1]. The random variable ρ has the distribution density nr^{n-1} on the segment [0, 1]. Finally, the random variable γ has a gamma distribution density $s^{n/2}\exp(-s)/\Gamma(1 + n/2)$.

If $R^2 < 2nt$, then the representation for $u(x, t)$ has a similar form:

$$I_1 = E\mathbf{1}_{\{\gamma \le \frac{n}{2\theta}\}}\frac{R^2}{2n} \cdot f\left(t - \frac{R^2}{2n}\theta, x + R\rho\sqrt{\frac{2\theta\gamma}{n}}\Omega\right), \tag{7}$$

$$I_2 = E\mathbf{1}_{\{\gamma > \frac{n}{2}\}}\mathbf{1}_{\left\{\theta > \frac{n}{2\gamma}\right\}}u\left(t - \frac{R^2}{4\gamma}, x + R\rho\Omega\right), \tag{8}$$

$$I_3 = E\mathbf{1}_{\{\gamma \le \frac{n}{2}\}}u\left(t - \frac{R^2}{2n}, x + R\rho\sqrt{\frac{2\gamma}{n}}\Omega\right), \tag{9}$$

$$I_4 = E\mathbf{1}_{\{\gamma > \frac{n}{2}\}}\mathbf{1}_{\left\{\theta \le \frac{n}{2\gamma}\right\}}u\left(t - \frac{R^2}{4\gamma}, x + R\Omega\right). \tag{10}$$

Now we define simulation procedure for RWC $\big(t(k), x(k)\big)$ and a sequence of random estimators ξ_k for the solution $u(t, x)$.

Let $\gamma_k, \rho_k, \theta_k, \Omega_k (k = 1, 2, \ldots)$ be the independent realizations of random variables defined previously. The distance $R(x(k))$ denote by R_k. Let $x(0) = x, t(0) = t$, $\xi_0 = u(t, x)$. It is obvious that the sum of random variables placed under the sign of the expectation in terms I_1, I_2, I_3, I_4 is an unbiased estimate for $u(t, x)$. But only one factor before the function u is nonzero. The arguments of function u in this random variable determine the next point for the RWC. For example, if $R_k^2 < 2nt(k)$ and $\gamma_k \le n/2$ then

$$t(k+1) = t(k) - \frac{R_k^2}{2n}, \quad x(k+1) = x(k) + R_k\rho_{k+1}\sqrt{\frac{2\gamma_{k+1}}{n}}\Omega_{k+1}.$$

A new estimator ξ_{k+1} is obtained by replacing the function $u(t(k), x(k))$ in the estimator ξ_k with its estimator by the formulas (7)–(10). So, under the selected conditions, the function $u(t(k), x(k))$ in the estimator ξ_k is replaced by the sum

$$u(t(k+1), x(k+1)) + \frac{R_k^2}{2n} \cdot f\left(t(k) - \frac{R_k^2}{2n}\theta_{k+1}, x(k) + R_k\rho_{k+1}\sqrt{\frac{2\theta_{k+1}\gamma_{k+1}}{n}}\Omega_{k+1}\right).$$

In the second case, the process and estimators are defined in a similar way. The RWC process stops at the moment N_1, such that $t(N_1) = 0$.

Some of the properties of the RWC process includes the following theorem (see [3]).

THEOREM 1.

1. *The sequence of estimators* $\xi_k \quad (k = 1, 2, \ldots)$ *is a square integrable martingale.*
2. *The RWC process* $\big(t(k), x(k)\big) \quad (k = 1, 2, \ldots)$ *converges to* $(t_\infty, x_\infty) \in \{0\} \times D \cup [0, t] \times \Gamma$ *with probability 1.*
3. *Let* $\delta > 0$, $N_2 = \min\{k : R(x(k)) < \delta\}$ *and* $N_\delta = \min(N_1, N_2)$. *Then estimator* ξ_{N_δ} *is unbiased for* $u(t, x)$.
4. *Let* $\eta = \delta^2/\max(\gamma, n/2)$. *Then* $EN_\delta - 1 \le t/E\eta + o(t)$.

Remark 1. In practice, we need to correct the estimator ξ_{N_δ}. Indeed, the value of $\xi_{N_2} = u(t(N_2), x(N_2))$ is unknown for $N_\delta = N_2$. But we can change this value by $\psi(t(N_2), x^*(N_2))$, where $x^*(N_2) \in \Gamma$ and $\|x^*(N_2) - x(N_2)\| < \delta$. A new estimator denote by $\xi^*_{N_\delta}$.

3 Random Walk on the Boundary Algorithm

Now we describe one of the variants of the Random Walk on the boundary algorithm (Sabelfeld, Simonov [4]). For simplicity, we assume the domain D to be convex. We write the solution of problem (1) as the sum of thermal potentials:

$$u(t, x) = u_1(t, x) + u_2(t, x) + u_3(t, x), \tag{11}$$

where

$$u_1(t, x) = t \cdot E\mathbf{1}_{\left\{\theta\gamma \le \frac{r^2}{4t}\right\}} f\left(t - t\theta, x + 2\sqrt{t\theta\gamma}\Omega\right), \tag{12}$$

$$u_2(t, x) = E\mathbf{1}_{\left\{\gamma \le \frac{r^2}{4t}\right\}} \varphi(x + 2\sqrt{t\gamma}\Omega), \tag{13}$$

where r is a distance between the points x and $y \in \varGamma$, which is visible from point x in direction $\varOmega$. The random variable γ has a density $p_n(s) = s^{n/2-1}\exp(-s)/\Gamma(n/2)$. The last function is a solution to a boundary value problem

$$\begin{aligned} \frac{\partial u_3(t,x)}{\partial t} &= \Delta u_3(t,x), \quad x \in D, \quad t > 0 \\ u_3(t,x) = \psi(t,x) - u_1(t,x) - u_2(t,x), & \quad x \in \varGamma, \quad t > 0 \\ u_3(0,x) &= 0, \quad x \in D. \end{aligned} \tag{14}$$

The function $u_3(t,x)$ is a double-layer potential

$$u_3(t,x) = E\mathbf{1}_{\left\{\gamma > \frac{r^2}{4t}\right\}} \mu\left(t - \frac{r^2}{4\gamma}, x + r\varOmega\right), \tag{15}$$

where $y = x + r\varOmega$ is a point of surface $\varGamma$, which is visible from point x in direction $\varOmega$.

The density $\mu(t,x)$ satisfies the integral equation for $x \in \varGamma$

$$\mu(t,x) = -E\mathbf{1}_{\left\{\gamma > \frac{r^2}{4t}\right\}} \mu\left(t - \frac{r^2}{4\gamma}, x + r\varOmega\right) + 2g(t,x), \tag{16}$$

where $g(t,x) = \psi(t,x) - u_1(t,x) - u_2(t,x)$ for $t > 0$.

Remark 2. Note that in this equation vector $\varOmega$ is uniformly distributed in the hemisphere defined by the inequality $(\varOmega, \nu_x) < 0$, where ν_x is an external normal vector to the surface at point $x \in \varGamma$.

Now we define simulation procedure for RWB $\big(t(k), x(k)\big)$ and a random estimator ξ for the solution $u(t,x)$.

Let $x \in D,\ x(0) = x,\ t(0) = t > 0,\ \xi = 0$ and $\gamma_k, \theta_k, \varOmega_k (k = 1, 2, \ldots)$ be the independent realizations of random variables with a gamma distribution (with the density $p_n(s)$) and a uniform distribution on [0, 1] and uniform distribution on unit sphere respectively.

At the first step,

1. We calculate the point $y \in \varGamma$ at which the ray $x(0) + r\varOmega_1$, $r > 0$ intersects the surface $\varGamma$;
2. If the condition $\theta_1\gamma_1 < r^2/4t$ is fulfilled, then $\xi := \xi + t \cdot f\,(t - t\theta_1, x + 2\sqrt{t\theta_1\gamma_1}\varOmega_1\big)$;
3. If the condition $\gamma_1 > r^2/4t$ is fulfilled, then we define a new point $x(1) = x + r\varOmega_1$ and a new time $t(1) = t - r^2/4\gamma_1$ and $Q := 1$;
4. If the condition $\gamma_1 \le r^2/4t$ is fulfilled, then $\xi := \xi + \varphi(x + 2\sqrt{t\gamma_1}\varOmega_1)$ and the construction of the estimator is completed.

Similar actions are performed in step k:

1. If $(\Omega_k, \nu_{x(k-1)}) > 0$, then $\Omega_k := -\Omega_k$;
2. $\xi := \xi + Q \cdot 2\psi(t(k-1), x(k-1))$;
3. We calculate the point $y \in \Gamma$ at which the ray $x(k-1) + r\Omega_k$, $r > 0$ intersects the surface Γ;
4. If the condition $\theta_k \gamma_k < r^2/4t(k-1)$ is fulfilled, then $\xi := \xi - Q \cdot t \cdot f\left(t(k-1) - t(k-1)\theta_k, x(k-1) + 2\sqrt{t(k-1)\theta_k\gamma_k}\Omega_k\right)$;
5. If the condition $\gamma_k > r^2/4t(k-1)$ is fulfilled, then we define new point $x(k) = x(k-1) + r\Omega_k$ and new time $t(k) = t(k-1) - r^2/4\gamma_k$ and $Q := -Q$;
6. If the condition $\gamma_k \le r^2/4t(k-1)$ is fulfilled, then $\xi := \xi - Q \cdot \varphi(x(k-1) + 2\sqrt{t(k-1)\gamma_k}\Omega_k)$ and the construction of the estimator is completed.

Remark 3. We constructed an estimator, which should be called the absorption-collision estimator. It is the collision estimator for $\psi(t, x)$ function and absorption estimator for $\varphi(t, x)$ function. Other estimators and process properties can be found in [2, 4].

4 Numerical Results

Consider the boundary problem for the ball $\| x \| \le \rho$ in R^3. To determine the point of intersection of the sphere with the ray, we obtain the formula

$$r = -(\Omega, x) + \sqrt{(\Omega, x)^2 + \rho^2 - \| x \|^2}, \tag{17}$$

in case $\| x \| < \rho$ and the formula

$$r = -2(\Omega, x), \tag{18}$$

in case $\| x \| = \rho$.

Let $n = 3$, $\rho = 1$ and $u(t, x) = \exp(-t) \cdot \sin(x_1 + t)$. Then

$$f(t, x) = \exp(-t) \cdot \cos(x_1 + t).$$

We calculate the values of the function $u(t, p)$ for a point $p = (0.5, 0.5, 0.5)$ for different values of time. All results are shown for computing on a single core AMD Ryzen 7 2700 processor. Parameter $\delta = 10^{-8}$ for the RWC process. The number of trajectories is denoted by M. The variable L denotes the average length of the trajectory, and the variable T stands for processor time.

Table 1 Values of the function $u(t, p) = \exp(-t) \cdot \sin(x_1 + t)$ at the point $p = (0.5, 0.5, 0.5)$. Monte Carlo method for (RWC) and (RWB) processes. ($t = 0.5$, $u(t, p) = 0.5103779515$)

M	RWC	ERR	RWB	ERR	L RWC	L RWB	T RWC	T RWB
10^4	0.51132	0.00419	0.51135	0.01811	87	3		
10^5	0.51096	0.00134	0.50978	0.00574	87	3		
10^6	0.51033	0.00042	0.51073	0.00182	87	3		
10^7	0.51034	0.00013	0.51025	0.00057	87	3	177 c	
10^8			0.51037	0.00018		3		53 c

Table 2 Values of the function $u(t, p) = \exp(-t) \cdot \sin(x_1 + t)$ at the point $p = (0.5, 0.5, 0.5)$. Monte Carlo method for (RWC) and (RWB) processes. ($t = 5$, $u(t, p) = -0.004753893319$)

M	RWC	ERR	RWB	ERR	L RWC	L RWB	T RWC	T RWB
10^4	−0.0047598	0.0000346	−0.010500	0.024800	87	12		
10^5	−0.0047580	0.0000109	−0.004316	0.007953	87	12		
10^6	−0.0047566	0.0000035	−0.003726	0.002516	87	12		
10^7	−0.0047547	0.0000011	−0.004548	0.000796	87	12	177 c	
10^8			−0.004742	0.000252		12		201 c

Table 3 Values of the function $u(t, p) = \exp(-t) \cdot \sin(x_1 + t)$ at the point $p = (0.5, 0.5, 0.5)$. Monte Carlo method for (RWC) and (RWB) processes. ($t = 10$, $u(t, p) = -0.00003993812572$)

M	RWC	ERR	RWB	ERR	L RWC	L RWB	T RWC	T RWB
10^4	−0.0000399551	0.0000003062	−0.0738000	0.02800	87	21		
10^5	−0.0000399698	0.0000000969	−0.0733000	0.00887	87	21		
10^6	−0.0000399287	0.0000000305	−0.0000911	0.00279	87	21		
10^7	−0.0000399376	0.0000000096	−0.0000348	0.00089	87	21	185 c	
10^8			−0.0000302	0.00028		21		345 c

Table 4 Values of the function $u(t, p) = \exp(-t) \cdot \sin(x_1 + 2 \cdot x_2 - x_3 + t)$ at the point $p = (0.5, 0.5, 0.5)$. Monte Carlo method for (RWC) and (RWB) processes. ($t = 3$, $u(t, p) = -0.03767897757$)

M	RWC	ERR	RWB	ERR	L RWC	L RWB	T RWC	T RWB
10^4	−0.037844	0.000885			87			
10^5	−0.037689	0.000282	−0.032908	0.031248	87	8		
10^6	−0.037690	0.000089	−0.031764	0.009878	87	8		
10^7	−0.037677	0.000028	−0.038145	0.003125	87	8	175 c	
10^8			−0.037507	0.000988		8		
10^9			−0.037628	0.000313		8		24 m. 02 c

Table 5 Values of the function $u(t, p) = t$ at the point $p = (0.5, 0.5, 0.5)$. Monte Carlo method for (RWC) and (RWB) processes. ($t = 10$, $u(t, p) = 10$)

M	RWC	ERR	RWB	ERR	L RWC	L RWB	T RWC	T RWB
10^4	10.000029	0.000448	9.914594	0.187085	87			
10^5	10.000009	0.000138	9.941957	0.059305	87	21		
10^6	9.9999996	0.000043	9.991535	0.018762	87	21		
10^7	9.9999950	0.000014	9.999740	0.005936	87	21	164 c	
10^8			10.000043	0.001877		21		286 c

Table 6 Values of the function $u(t, p) = unknown$ at the point $p = (0.5, 0.5, 0.5)$. Monte Carlo method for (RWC) and (RWB) processes. ($t = 10$, $f(t, p) = 0$, $\varphi(t, p) = 0$, $\psi(t, p) = t$,)

M	RWC	ERR	RWB	ERR	L RWC	L RWB	T RWC	T RWB
10^4	9.95899	0.00203	9.98897	0.12767	87			
10^5	9.95862	0.00065	9.95060	0.04002	87	21		
10^6	9.95835	0.00021	9.95560	0.01263	87	21		
10^7	9.95834	0.00007	9.95791	0.00400	87	21	164 c	
10^8			9.95916	0.00126		21		281 c

5 Conclusion

The complexity of the algorithm for solving the problem was determined by the time of its operation to achieve the required error. Therefore, each table presents one run time for each algorithm. These times correspond to sample sizes at which the algorithms give the same error. Analyzing the results, it can be noted that

1. For “small” values of t, the RWB algorithm works better than algorithm RWC. (Table 1)
2. For “large” values of t, the RWC algorithm works better than algorithm RWB. (Tables 2, 3)
3. For the complicated function $u(t, x)$ the RWC algorithm works significantly better than the RWB algorithm. (Table 4)
4. The complexity of the RWB algorithm grows with increasing time even on a simple function $u(t, x)$. (Table 5)
5. If the potentials $u_1(t, x)$, $u_2(t, x)$ are non-zero, the variance of the estimators of the RWB algorithm increases quickly with increasing time. (Tables 3, 5)
6. If the potentials $u_1(t, x)$, $u_2(t, x)$ are zero, the variance of the estimators of the RWB 100 times more than the variance of the estimators of the RWC. (Table 6)

Perhaps the problems of the RWB algorithm are related to the fact that the estimators of the thermal potentials have a large variance. It would be interesting to find estimates of potentials with less variance.

A feature of the RWC algorithm is the fulfillment of the condition $\tau < R^2/(2n)$ for a time step τ similar to the stability condition for an explicit difference scheme. It is possible that the success of the RWC algorithm is associated with it.

Acknowledgements The research has been supported by the Russian Science Foundation under grant 19-11-00020.

References

1. Ermakov, S.M., Nekrutkin, V.V., Sipin, A.S.: Random Processes for Classical Equations of Mathematical Physics. Kluwer Academic Publishers, Dordrecht (1989)
2. Sabelfeld, K.K.: Monte Carlo Methods in Boundary Value Problems. Springer, Heidelberg (1991)
3. Ermakov, S.M., Sipin, A.S.: Monte-Carlo method and parametric separability of algorithms. Publishing House of St. Petersburg State University, St. Petersburg (2014). 247 p
4. Sabelfeld, K.K., Simonov, N.A.: Stochastic Methods for Boundary Value Problems. Numerics for High-Dimensional PDEs and Applications. Walter de Gruyter, Berlin (2016). 200 p
5. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**(3), 607–617 (2008)
6. Cliffe, K.A., Giles, M.B., Scheichl, R., Teckentrup, A.: Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. Comput. Vis. Sci. **14**(1), 315 (2011)
7. Sipin, A.S., Zeifman, A.I.: Continuous version of the Kellog algorithm and its Monte Carlo realization. In: AIP Conference Proceedings, vol. 1863, p. 090008 (2017)

Local Controllability of a Class of Fractional Differential Inclusions via Derived Cones

Aurelian Cernea

Abstract We study a class of fractional differential inclusions defined by Caputo-Katugampola fractional derivative and we provide a sufficient condition for local controllability along a reference trajectory. This condition is obtain in terms of a certain variational fractional differential inclusion associated to the problem studied. More exactly, we prove that the reachable set of a certain variational fractional differential inclusion of Caputo-Katugampola type is a derived cone in the sense of Hestenes to the reachable set of the problem and then, in order to obtain our main result, we essentially use an outstanding property of derived cones and a continuous version of Filippov's theorem for solutions of the fractional differential inclusion considered.

Keywords Differential inclusion · Fractional derivative · Derived cone · Local controllability

1 Introduction

The last years represent a period of strong development of the theory of differential equations and inclusions of fractional order ([3, 8, 12, 13, 16] etc.). This is justified by the fact that fractional differential equations are very useful tools in order to model many physical phenomena.

Recently, a generalized Caputo-Katugampola fractional derivative was proposed in [11] by Katugampola and afterwards he provided the existence of solutions for fractional differential equations defined by this derivative. This Caputo-Katugampola fractional derivative extends the well known Caputo and Caputo-Hadamard fractional derivatives into a single form. Even if Katugampola fractional integral operator is

A. Cernea (✉)
Faculty of Mathematics and Computer Science, University of Bucharest,
Academiei 14, 010014 Bucharest, Romania
e-mail: acernea@fmi.unibuc.ro

Academy of Romanian Scientists, Splaiul Independenţei 54, 050094 Bucharest, Romania

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_12

an Erdélyi-Kober type operator [9] it is argued [11] that is not possible to derive Hadamard equivalence operators from Erdélyi-Kober type operators. Also, in some recent papers [1, 7, 17], several qualitative properties of solutions of fractional differential equations defined by Caputo-Katugampola derivative were obtained.

In this paper we study the following problem

$$D_c^{\alpha,\rho}x(t) \in F(t, x(t)) \quad a.e.\ ([0, T]), \quad x(0) \in X_0, \quad x'(0) \in X_1, \tag{1}$$

where $\alpha \in (1, 2]$, $\rho \geq 1$, $D_c^{\alpha,\rho}$ is the Caputo-Katugampola fractional derivative, F: $[0, T] \times \mathbf{R} \to \mathscr{P}(\mathbf{R})$ is a set-valued map and X_0, $X_1 \subset \mathbf{R}$ are closed sets.

The aim of the present paper is to provide a sufficient condition for local controllability along a reference trajectory of differential inclusion (1) in terms of certain variational fractional differential inclusion associated to problem (1).

A key tool in our approach is the notion of derived cone to an arbitrary subset of a normed space introduced by M. Hestenes in [10]. Initially this concept was used in [10] to obtain necessary optimality conditions in control theory. Other properties of derived cones obtained in [14, 15] are useful to obtain several results in the qualitative theory of control systems.

We prove that the reachable set of a certain variational fractional differential inclusion of Caputo-Katugampola type is a derived cone in the sense of Hestenes to the reachable set of the problem (1). In order to obtain the continuity property in the definition of a derived cone we shall use a continuous version of Filippov's theorem for solutions of fractional differential inclusion (1) recently obtained in [7] (see also [4, 6]).

We note that a similar result for fractional differential inclusions defined by Caputo fractional derivative may be found in [5]; therefore, the present paper may be regarded as an extension of the results in [5] to the more general problem (1).

The paper is organized as follows: in Sect. 2 we present the notations and the preliminary results to be used in the sequel and in Sect. 3 we provide our main results.

2 Preliminaries

In general the reachable set to a control system is, generally, neither a differentiable manifold, nor a convex set, its infinitesimal properties may be characterized only by tangent cones in a generalized sense, extending the classical concepts of tangent cones in differential geometry and convex analysis, respectively.

Definition 1 [10] A subset $D \subset \mathbf{R}^n$ is said to be a *derived set to* $X \subset \mathbf{R}^n$ at $x \in X$ if for any finite subset $\{w_1, ..., w_k\} \subset D$, there exist $s_0 > 0$ and a continuous mapping $\alpha(.) : [0, s_0]^k \to X$ such that $\alpha(0) = x$ and $\alpha(.)$ is (conically) differentiable at $s = 0$ with the derivative col$[w_1, ..., w_k]$ in the sense that

$$\lim_{R_+^k \ni \theta \to 0} \frac{||\alpha(\theta) - \alpha(0) - \sum_{i=1}^k \theta_i w_i||}{||\theta||} = 0.$$

We shall write in this case that the derivative of $\alpha(.)$ at $s = 0$ is given by

$$D\alpha(0)\theta = \sum_{i=1}^{k} \theta_j w_j \quad \forall \theta = (\theta_1, ..., \theta_k) \in \mathbf{R}_+^k := [0, \infty)^k.$$

A subset $C \subset \mathbf{R}^n$ is said to be a *derived cone* of X at x if it is a derived set and also a convex cone.

For the basic properties of derived sets and cones we refer to M. Hestenes [10]; we recall that if D is a derived set then $D \bigcup \{0\}$ as well as the convex cone generated by D, defined by

$$cco(D) = \{\sum_{i=1}^{k} \lambda_j w_j; \quad \lambda_j \geq 0,\ k \in N,\ w_j \in D,\ j = 1, ..., k\}$$

is also a derived set, hence a derived cone.

The fact that the derived cone is a proper generalization of the classical concepts in differential geometry and convex analysis is illustrated by the following results [10]: if $X \subset \mathbf{R}^n$ is a differentiable manifold and $T_x X$ is the tangent space in the sense of differential geometry to X at x

$$T_x X = \{w \in \mathbf{R}^n;\ \exists\, c : (-s, s) \to \mathbf{R}^n,\ \text{of class } C^1, c(0) = x, c'(0) = w\},$$

then $T_x X$ is a derived cone; also, if $X \subset \mathbf{R}^n$ is a convex subset then the tangent cone in the sense of convex analysis defined by

$$TC_x X = cl\{t(y - x); \quad t \geq 0,\ y \in X\}$$

is also a derived cone. Since any convex subcone of a derived cone is also a derived cone, such an object may not be uniquely associated to a point $x \in X$; moreover, simple examples show that even a maximal with respect to set-inclusion derived cone may not be uniquely defined: if the set $X \subset \mathbf{R}^2$ is defined by

$$X = C_1 \bigcup C_2, \quad C_1 = \{(x, x); x \geq 0\}, \quad C_2 = \{(x, -x), x \leq 0\},$$

then C_1 and C_2 are both maximal derived cones of X at the point $(0, 0) \in X$.

At the same time, the up-to-date experience in nonsmooth analysis shows that for some problems, the use of one of the intrinsic tangent cones may be preferable. The most known intrinsic tangent cones in the literature (e.g. [2]) are the contingent, the quasitangent (intermediate) and Clarke's tangent cones, defined, respectively, by

$$K_xX = \{v \in X; \quad \exists s_m \to 0+, \quad \exists x_m \to x, \;\; x_m \in X: \;\; \frac{x_m - x}{s_m} \to v\},$$

$$Q_xX = \{v \in X; \quad \forall s_m \to 0+, \quad \exists x_m \to x, \;\; x_m \in X: \;\; \frac{x_m - x}{s_m} \to v\},$$

$$C_xX = \{v \in X; \quad \forall (x_m, s_m) \to (x, 0+), \quad x_m \in X, \; \exists y_m \in X: \;\; \frac{y_m - x_m}{s_m} \to v\}.$$

The next property of derived cone, obtained by Hestenes ([10], Theorem 4.7.4) and stated in the next lemma is essential in the proof of our main result.

Lemma 1 *Let $X \subset \mathbf{R}^n$. Then $x \in int(X)$ if and only if $C = \mathbf{R}^n$ is a derived cone at $x \in X$ to X.*

Corresponding to each type of tangent cone, say $\tau_x X$ one may introduce (e.g. [2]) a *set-valued directional derivative* of a multifunction $G(.) : X \subset \mathbf{R}^n \to \mathscr{P}(\mathbf{R}^n)$ (in particular of a single-valued mapping) at a point $(x, y) \in Graph(G)$ as follows

$$\tau_y G(x; v) = \{w \in \mathbf{R}^n; \quad (v, w) \in \tau_{(x,y)} Graph(G)\}, \quad v \in \tau_x E.$$

We recall that a set-valued map, $A(.) : \mathbf{R}^n \to \mathscr{P}(\mathbf{R}^n)$ is said to be a *convex* (respectively, closed convex) *process* if $Graph(A(.)) \subset \mathbf{R}^n \times \mathbf{R}^n$ is a convex (respectively, closed convex) cone. For the basic properties of convex processes we refer to [2], but we shall use here only the above definition.

Let $T > 0$, $I := [0, T]$ and denote by $\mathscr{L}(I)$ the σ-algebra of all Lebesgue measurable subsets of I. Denote by $\mathscr{P}(\mathbf{R})$ the family of all nonempty subsets of $\mathbf{R}$ and by $\mathscr{B}(\mathbf{R})$ the family of all Borel subsets of $\mathbf{R}$.

As usual, we denote by $C(I, \mathbf{R})$ the Banach space of all continuous functions $x(.) : I \to \mathbf{R}$ endowed with the norm $|x(.)|_C = \sup_{t \in I} |x(t)|$ and by $L^1(I, \mathbf{R})$ the Banach space of all (Bochner) integrable functions $x(.) : I \to \mathbf{R}$ endowed with the norm $|x(.)|_1 = \int_0^T |x(t)| dt$.

In [11] the following notions were introduced. Let $\rho > 0$.

Definition 2 ([11]) (a) *The generalized left-sided fractional integral of order $\alpha > 0$* of a Lebesgue integrable function $f : [0, \infty) \to \mathbf{R}$ is defined by

$$I^{\alpha,\rho} f(t) = \frac{\rho^{1-\alpha}}{\Gamma(\alpha)} \int_0^t (t^\rho - s^\rho)^{\alpha-1} s^{\rho-1} f(s) ds,$$

provided the right-hand side is pointwise defined on $(0, \infty)$ and $\Gamma(.)$ is (Euler's) Gamma function defined by $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.

(b) *The generalized fractional derivative*, corresponding to the generalized left-sided fractional integral of a function $f : [0, \infty) \to \mathbf{R}$ is defined by

$$D^{\alpha,\rho} f(t) = (t^{1-\rho} \frac{d}{dt})^n (I^{n-\alpha,\rho})(t) = \frac{\rho^{\alpha-n+1}}{\Gamma(n-\alpha)} (t^{1-\rho} \frac{d}{dt})^n \int_0^t \frac{s^{\rho-1} f(s)}{(t^\rho - s^\rho)^{\alpha-n+1}} ds$$

if the integral exists and $n = [\alpha] + 1$.

(c) *The Caputo-Katugampola* generalized fractional derivative is defined by

$$D_c^{\alpha,\rho} f(t) = (D^{\alpha,\rho}[f(s) - \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!} s^k])(t)$$

We note that if $\rho = 1$, the Caputo-Katugampola fractional derivative becames the well known Caputo fractional derivative. On the other hand, passing to the limit with $\rho \to 0+$, the above definition yields the Hadamard fractional derivative.

Definition 3. A function $x(.) \in C(I, \mathbf{R})$ is called a solution of problem (1) if there exists a function $v(.) \in L^1(I, \mathbf{R})$ with $v(t) \in F(t, x(t))$, $a.e.$ (I) such that $D_c^{\alpha,\rho} x(t) = v(t)$, $a.e.$ (I) and $x(0) = x_0 \in X_0$, $x'(0) = x_1 \in X_1$.

In this case we say that $(x(.), v(.))$ is a *trajectory-selection pair* of (1).

Hypothesis 1 (i) *$F(.,.) : I \times \mathbf{R} \to \mathscr{P}(\mathbf{R})$ has nonempty closed values and is $\mathscr{L}(I) \otimes \mathscr{B}(\mathbf{R})$ measurable.*

(ii) *There exists $L(.) \in L^1(I, (0, \infty))$ such that, for almost all $t \in I$, $F(t, .)$ is $L(t)$-Lipschitz in the sense that*

$$d_H(F(t, x), F(t, y)) \leq L(t)|x - y| \quad \forall x, y \in \mathbf{R},$$

where $d_H(.,.)$ is the Hausdorff distance

$$d(A, B) = \max\{d^*(A, B), d^*(B, A)\}, \quad d^*(A, B) = \sup\{d(a, B); a \in A\}.$$

Hypothesis 2 (i) *S is a separable metric space, $a(.), b(.) : S \to \mathbf{R}$ and $c(.) : S \to (0, \infty)$ are continuous mappings.*

(ii) *There exists the continuous mappings $g(.), p(.) : S \to L^1(I, \mathbf{R})$, $y(.) : S \to C(I, \mathbf{R})$ such that*

$$(Dy(s))_c^{\alpha,\rho}(t) = g(s)(t) \quad a.e.\ t \in I, \quad \forall s \in S,$$

$$d(g(s)(t), F(t, y(s)(t)) \leq p(s)(t) \quad a.e.\ t \in I,\ \forall\ s \in S.$$

We use next the following notation

$$\xi(s) = \frac{1}{1 - |I^{\alpha,\rho} L|}(|a(s) - y(s)(0)| + T|b(s) - (y(s))'(0)| + c(s) + |I^{\alpha,\rho} p(s)|),\ s \in S,$$

where $|I^{\alpha,\rho} L| := \sup_{t \in I} |I^{\alpha,\rho} L(t)|$ and $|I^{\alpha,\rho} p(s)| := \sup_{t \in I} |I^{\alpha,\rho} p(s)(t)|$.

The main tool in characterizing derived cones to reachable sets of fractional differential inclusions is a certain version of Filippov's theorem for fractional differential inclusion (1) in [7].

Theorem 1 ([7]) *Assume that Hypotheses 1 and 2 are satisfied.*

If $|I^{\alpha,\rho}L| < 1$, *then there exist a continuous mapping* $x(.) : S \to C(I, \mathbf{R})$ *such that for any* $s \in S$, $x(s)(.)$ *is a solution of problem*

$$D_c^{\alpha,\rho} z(t) \in F(t, z(t)), \quad z(0) = a(s), \quad z'(0) = b(s)$$

such that

$$|x(s)(t) - y(s)(t)| \leq \xi(s) \quad \forall (t, s) \in I \times S. \tag{2}$$

3 The Main Results

We study next the reachable set of (1) defined by

$$R_F(T, X_0, X_1) := \{x(T); \quad x(.) \text{ is a solution of (1)}\}.$$

We consider a certain variational fractional differential inclusion and we shall prove that the reachable set of this variational inclusion from derived cones $C_0 \subset \mathbf{R}$ to X_0 and $C_1 \subset \mathbf{R}$ to X_1 at time T is a derived cone to the reachable set $R_F(T, X_0, X_1)$.

Throughout in this section we assume the folowwing hypotheses.

Hypothesis 3 (i) *Hypothesis 1 is satisfied,* $\alpha \in (1, 2]$, $\rho \geq 1$, $|I^{\alpha,\rho}L| < 1$ *and* $X_0, X_1 \subset \mathbf{R}$ *are closed sets.*

(ii) $(z(.), f(.)) \in C(I, \mathbf{R}) \times L^1(I, \mathbf{R})$ *is a trajectory-selection pair of (1) and a family* $A(t, .) : \mathbf{R} \to \mathscr{P}(\mathbf{R})$, $t \in I$ *of convex processes satisfying the condition*

$$A(t, u) \subset Q_{f(t)} F(t, .)(z(t); u) \quad \forall u \in dom(A(t, .)), \ a.e. \ t \in I \tag{3}$$

is assumed to be given and defines the variational inclusion

$$D_c^{\alpha,\rho} w(t) \in A(t, w(t)). \tag{4}$$

Remark 1 We mention that for any set-valued map $F(., .)$, one may find an infinite number of families of convex process $A(t, .)$, $t \in I$, satisfying condition (3); in fact any family of closed convex subcones of the quasitangent cones, $\overline{A}(t) \subset Q_{(z(t), f(t))} graph(F(t, .))$, defines the family of closed convex process

$$A(t, u) = \{v \in \mathbf{R}; \ (u, v) \in \overline{A}(t)\}, \quad u, v \in \mathbf{R}, \ t \in I$$

that satisfy condition (3). For example, we may take an "intrinsic" family of such closed convex process; namely, Clarke's convex-valued directional derivatives $C_{f(t)}F(t,.)(z(t);.)$.

When $F(t,.)$ is assumed to be Lipschitz a.e. on I an alternative characterization of the quasitangent directional derivative is (e.g., [2])

$$Q_{f(t)}F(t,.)((z(t);u)) = \{w \in \mathbf{R};\ \lim_{\theta \to 0+} \frac{1}{\theta} d(f(t)+\theta w, F(t, z(t)+\theta u)) = 0\}. \tag{5}$$

Theorem 2. *Assume that Hypothesis 3 is satisfied, let $C_0 \subset \mathbf{R}$ be a derived cone to X_0 at $z(0)$ and $C_1 \subset \mathbf{R}$ be a derived cone to X_1 at $z'(0)$. Then the reachable set $R_A(T, C_0, C_1)$ of (4) is a derived cone to $R_F(T, X_0, X_1)$ at $z(T)$.*

Proof In view of Definition 1, let $\{w_1, ..., w_m\} \subset R_A(T, C_0, C_1)$, hence such that there exist the trajectory-selection pairs $(v_1(.), g_1(.)), ..., (v_m(.), g_m(.))$ of the variational inclusion (4) such that

$$v_j(T) = w_j, \quad v_j(0) \in C_0, \quad v_j'(0) \in C_1, \quad j = 1, 2, ..., m \tag{6}$$

Since $C_0 \subset \mathbf{R}$ is a derived cone to X_0 at $z(0)$ and $C_1 \subset \mathbf{R}$ is a derived cone to X_1 at $z'(0)$, there exist the continuous mappings $\alpha_0 : S = [0, \theta_0]^m \to X_0, \alpha_1 : S \to X_1$ such that

$$\begin{array}{ll} \alpha_0(0) = z(0), & D\alpha_0(0)s = \sum_{j=1}^m s_j v_j(0) \quad \forall s \in \mathbf{R}_+^m, \\ \alpha_1(0) = z'(0), & D\alpha_1(0)s = \sum_{j=1}^m s_j v_j'(0) \quad \forall s \in \mathbf{R}_+^m. \end{array} \tag{7}$$

For any $s = (s_1, ..., s_m) \in S$ and $t \in I$ we set

$$\begin{array}{c} y(s)(t) = z(t) + \sum_{j=1}^m s_j v_j(t), \\ g(s)(t) = f(t) + \sum_{j=1}^m s_j g_j(t), \\ p(s)(t) = d(g(s)(t), F(t, y(s)(t))) \end{array} \tag{8}$$

and we prove that $y(.)$, $p(.)$ satisfy the hypothesis of Theorem 1.

From the lipschitzianity of $F(t,.)$ we have that for any $s \in S$, the measurable function $p(s)(.)$ in (8) it is also integrable.

$$p(s)(t) = d(g(s)(t), F(t, y(s)(t))) \leq \sum_{j=1}^m s_j |g_j(t)|$$

$$+ d_H(F(t, z(t)), F(t, y(s)(t))) \leq \sum_{j=1}^m s_j |g_j(t)| + L(t) \sum_{j=1}^m s_j |v_j(t)|.$$

At the same time, the mapping $s \to p(s)(.) \in L^1(I, \mathbf{R})$ is Lipschitzian (and, in particular, continuous) since for any $s, s' \in S$ one may write

$$|p(s)(.) - p(s')(.)|_1 = \int_0^T |p(s)(t) - p(s')(t)|dt \leq \int_0^T [|g(s)(t) - g(s')(t)| +$$

$$d_H(F(t, y(s)(t)), F(t, y(s')(t))))]dt \leq ||s - s'||(\sum_{j=1}^m \int_0^T [|g_j(t)| + L(t)|v_j(t)|]dt)$$

Define $S_1 := S\backslash\{(0, \ldots, 0)\}$ and $c(.) : S_1 \to (0, \infty)$, $c(s) := ||s||^2$. It follows from Theorem 1 the existence of a continuous function $x(.) : S_1 \to C(I, \mathbf{R})$ such that for any $s \in S_1$, $x(s)(.)$ is a solution of (1.1) with the property (2).

For $s = 0$ we define $x(0)(t) = y(0)(t) = z(t)$ $\forall t \in I$. Obviously, $x(.) : S \to C(I, \mathbf{R})$ is also continuous.

Finally, we define the function $\alpha(.) : S \to R_F(T, X_0, X_1)$ by

$$\alpha(s) = x(s)(T) \quad \forall s \in S.$$

Obviously, $\alpha(.)$ is continuous on S and verifies $\alpha(0) = z(T)$.

In order to finish the proof we must show that $\alpha(.)$ is differentiable at $s_0 = 0 \in S$ and its derivative is given by

$$D\alpha(0)(s) = \sum_{j=1}^m s_j w_j \quad \forall s \in \mathbf{R}_+^m$$

which is equivalent with the fact that:

$$\lim_{s \to 0} \frac{1}{||s||}(|\alpha(s) - \alpha(0) - \sum_{j=1}^m s_j w_j|) = 0. \tag{9}$$

Taking into account (2) we obtain

$$\frac{1}{||s||}|\alpha(s) - \alpha(0) - \sum_{j=1}^m s_j w_j| \leq \frac{1}{||s||}|x(s)(T) - y(s)(T)| \leq \frac{1}{1 - |I^{\alpha,\rho}L|}||s||$$
$$+ \frac{1}{1 - |I^{\alpha,\rho}L|}\frac{1}{||s||}|\alpha_0(s) - z(0) - \sum_{j=1}^m s_j v_j(0)| + \frac{T}{1 - |I^{\alpha,\rho}L|} \cdot$$
$$\frac{1}{||s||}|\alpha_1(s) - z'(0) - \sum_{j=1}^m s_j v_j'(0)| + \frac{T^{\rho\alpha-1}}{(1 - |I^{\alpha,\rho}L|)\Gamma(\alpha)\rho^{\alpha-1}} \int_0^T \frac{p(s)(u)}{||s||}du$$

and therefore in view of (7), relation (9) is implied by the following property of the mapping $p(.)$ in (8)

$$\lim_{s \to 0} \frac{p(s)(t)}{||s||} = 0 \quad a.e. (I). \tag{10}$$

In order to prove the last property we note since $A(t,.)$ is a convex process for any $s \in S$ one has

$$\sum_{j=1}^{m} \frac{s_j}{||s||} g_j(t) \in A(t, \sum_{j=1}^{m} \frac{s_j}{||s||} u_j(t)) \subset Q_{f(t)} F(t,.)(z(t); \sum_{j=1}^{m} \frac{s_j}{||s||} u_j(t)) \quad a.e.\ (I).$$

Therefore, by (5) we obtain

$$\lim_{h \to 0+} \frac{1}{h} d(f(t) + h \sum_{j=1}^{m} \frac{s_j}{||s||} g_j(t), F(t, z(t) + h \sum_{j=1}^{m} \frac{s_j}{||s||} v_j(t))) = 0. \tag{11}$$

Finally, in order to prove that (11) implies (10) we take the compact metric space $\Sigma_+^{m-1} = \{\sigma \in \mathbf{R}_+^m; ||\sigma|| = 1\}$ and the real function $\psi_t(.,.) : (0, \theta_0] \times \Sigma_+^{m-1} \to \mathbf{R}_+$ defined by

$$\psi_t(h, \sigma) = \frac{1}{h} d(f(t) + h \sum_{j=1}^{m} \sigma_j g_j(t), F(t, z(t) + h \sum_{j=1}^{m} \sigma_j v_j(t))), \tag{12}$$

where $\sigma = (\sigma_1, ..., \sigma_m)$ and which according to (11) has the property

$$\lim_{\theta \to 0+} \psi_t(\theta, \sigma) = 0 \quad \forall \sigma \in \Sigma_+^{m-1}\ a.e.\ (I) \tag{13}$$

Using the fact that $\psi_t(\theta, .)$ is Lipschitzian and the fact that Σ_+^{m-1} is a compact metric space, from (13) it follows easily that

$$\lim_{\theta \to 0+} \max_{\sigma \in \Sigma_+^{m-1}} \psi_t(\theta, \sigma) = 0$$

which implies the fact that

$$\lim_{s \to 0} \psi_t(||s||, \frac{s}{||s||}) = 0 \quad a.e.\ (I)$$

and the proof is complete. □

We apply now Theorem 2 in order to obtain a sufficient condition for local controllability of the fractional differential inclusion (1) along a reference trajectory, $z(.)$ at time T, in the sense that

$$z(T) \in Int(R_F(T, X_0, X_1)).$$

Theorem 3 *Let $z(.)$, $F(.,.)$ and $A(.,.)$ satisfy Hypothesis 3, let $C_0 \subset \mathbf{R}$ be a derived cone to X_0 at $z(0)$ and $C_1 \subset \mathbf{R}$ be a derived cone to X_1 at $z'(0)$. If, the* *variational fractional differential inclusion in (4) is controllable at T in the sense that $R_A(T, C_0, C_1) = \mathbf{R}$, then the differential inclusion (1) is locally controllable along $z(.)$ at time T.*

Proof The proof follows from Lemma 1 and Theorem 2. □

References

1. Almeida, R., Malinowski, A.B., Odzijewicz, T.: Fractional differential equations with dependence on the Caputo-Katugampola derivative. J. Comput. Nonlin. Dyn. **11**, 1–11 (2016). ID 061017
2. Aubin, J.P., Frankowska, H.: Set-Valued Analysis. Birkhauser, Basel (1990)
3. Băleanu, D., Diethelm, K., Scalas, E., Trujillo, J.J.: Fractional Calculus Models and Numerical Methods. World Scientific, Singapore (2012)
4. Cernea, A.: Continuous version of Filippov's theorem for fractional differential inclusions. Nonlinear Anal. **72**, 204–208 (2010)
5. Cernea, A.: Derived cones to reachable sets of fractional differential inclusions. Commun. Appl. Nonlin. Anal. **19**, 23–31 (2012)
6. Cernea, A.: Continuous selections of solutions sets of fractional integrodifferential inclusions. Acta Math. Sci. **35B**, 399–406 (2015)
7. Cernea, A.: Continuous family of solutions for fractional integro-differential inclusions of Caputo-Katugampola type. Progress Fract. Diff. Appl. **5**, 37–42 (2019)
8. Diethelm, K.: The Analysis of Fractional Differential Equations. Springer, Berlin (2010)
9. Erdélyi, A., Kober, H.: Some remarks on Hankel transforms. Quart. J. Math. **11**, 212–221 (1940)
10. Hestenes, M.R.: Calculus of Variations and Optimal Control Theory. Wiley, New York (1966)
11. Katugampola, U.N.: A new approach to generalized fractional derivative. Bull. Math. Anal. Appl. **6**, 1–15 (2014)
12. Kilbas, A., Srivastava, H.M., Trujillo, J.J.: Theory and Applications of Fractional Differential Equations. Elsevier, Amsterdam (2006)
13. Miller, K., Ross, B.: An Introduction to the Fractional Calculus and Differential Equations. Wiley, New York (1993)
14. Mirică, Ş.: New proof and some generalizations of the Minimum Principle in optimal control. J. Optim. Theory Appl. **74**, 487–508 (1992)
15. Mirică, Ş.: Intersection properties of tangent cones and generalized multiplier rules. Optimization **46**, 135–163 (1999)
16. Podlubny, I.: Fractional Differential Equations. Academic Press, San Diego (1999)
17. Zeng, S., Băleanu, D., Bai, Y., Wu, G.: Fractional differential equations of Caputo-Katugampola type and numerical solutions. Appl. Math. Comput. **315**, 549–554 (2017)

A Modified Second-Order Collatz Equation as a Mathematical Model of Bipolar Disorder

Candace M. Kent

Abstract We propose, for the sake of dialogue, that the following system of difference equations serve as a phenomenological model of bipolar disorder, a psychiatric illness characterized by cycles or recurrent episodes of severe disturbances in mood (i.e., in being happy or sad, emotions at opposite poles of the spectrum):

$$\begin{cases} x_{n+1} = (ax_n + b) \bmod m, \\ \\ z_{n+1} = \begin{cases} \dfrac{-z_n - z_{n-1}}{2}, & \text{if } z_n + z_{n-1} \text{ is even,} \\ -z_n - z_{n-1}, & \text{if } z_n + z_{n-1} \text{ is odd} \end{cases} + s\delta(x_n), \end{cases}$$

where $a, b, m \in \mathbf{N}$, $x_0 \in \mathbf{N} \cup \{0\}$, $s \in \mathbf{Z} - \{0\}$, $z_{-1}, z_0 \in \mathbf{Z}$, and

$$\delta(x) = \begin{cases} 0, \text{ if } x \neq d \in \{0, 1, \ldots, m-1\}, \\ \\ 1, \text{ if } x = d. \end{cases}$$

The first equation in the system is a *linear congruential sequence*, used to generate a pseudo-random sequence of numbers; and the second equation is a modified (with the addition of $s\delta(x_n)$) version of one of the sixteen *Collatz difference equations* investigated by Amleh and colleagues in 1998. Let c be an odd scalar. While every solution of the (unmodified) Collatz equation is eventually periodic in the three-cycle $(c, 0, -c)$ or $(-c, 0, c)$, we observe (and conjecture) that every solution $\{z_n\}_{n=0}^{\infty}$ of the system above is also eventually periodic, but not in the three-cycle $(c, 0, -c)$ or $(-c, 0, c)$ and instead contains infinitely many recurrences of the interrupted three-cycle $(c, 0, -c)$ or $(-c, 0, c)$. Thus, a solution $\{z_n\}_{n=0}^{\infty}$ of the system is intended to represent the recurrent episodes of mood disturbance seen in an individual with bipolar disorder.

C. M. Kent (✉)
Department of Mathematics & Applied Mathematics, Virginia Commonwealth University,
1015 Floyd Avenue, P.O. Box 842014, Richmond, VA 23284-2014, USA
e-mail: cmkent@vcu.edu

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_13

1 Introduction and Preliminaries

Bipolar disorder, in the distant past referred to as manic-depressive psychosis, is a psychiatric illness that has a lifetime prevalence (i.e., proportion of the population that has had a disease at some point during a lifetime [5]) of 1.2% and that is characterized by extreme disturbances in mood. This disorder is considered cyclical in nature, where an individual suffering from it experiences recurrent episodes of what is called *mania* (i.e., excessive euphoria and unusually high energy) or *major depression* (i.e., profound sadness and low energy that is incapacitating), but with at least one of the episodes being that of mania. These recurrent episodes of mania or depression usually alternate with periods of remission in which the bipolar individual is what is called *euthymic*, i.e., (temporarily) healthy and in a state of well-being.

The euphoric mood in mania is associated with features which fall under the following three categories: (1) hyperactivity; (2) increased risk-taking; and (3) increased pleasure-seeking behavior [5]. The sadness in major depression has the associated features of loss of appetite or overeating; insomnia or hypersomnia (i.e., too much sleeping); psychomotor retardation (i.e., a slowing down of movement and thinking) or psychomotor agitation (i.e., agitated movement and thinking); feelings of worthlessness or guilt; and a loss of pleasure in all or almost all activities [3].

Many individuals with bipolar disorder can be successfully treated with medication, called *mood-stabilizing drugs*, such as lithium carbonate or anticonvulsant drugs. However, there is no cure for the illness, and a bipolar individual usually must remain on medication for a lifetime.

We propose, for the sake of dialogue, that the following system of difference equations serve as a phenomenological model of the longitudinal (i.e., lifetime) course of *untreated* bipolar disorder:

$$\begin{cases} x_{n+1} = (ax_n + b) \bmod m, \quad n = 0, 1, \ldots, \\ \\ z_{n+1} = \begin{cases} \dfrac{-z_n - z_{n-1}}{2}, & \text{if } z_n + z_{n-1} \text{ is even,} \\ -z_n - z_{n-1}, & \text{if } z_n + z_{n-1} \text{ is odd} \end{cases} + s\delta(x_n), \quad n = 0, 1, \ldots, \end{cases} \tag{1}$$

where $a, b, m \in \mathbf{N}$, $x_0 \in \mathbf{N} \cup \{0\}$, $s \in \mathbf{Z} - \{0\}$, $z_{-1}, z_0 \in \mathbf{Z}$, and

$$\delta(x) = \begin{cases} 0, & \text{if } x \neq d \in \{0, 1, \ldots, m-1\}, \\ \\ 1, & \text{if } x = d. \end{cases}$$

The first equation in system (1) that forms our mathematical model,

$$\begin{aligned} &x_{n+1} = (ax_n + b) \bmod m, \quad n = 0, 1, \ldots, \\ &x_0 \in \mathbf{N} \cup \{0\}, \quad a, b \in \mathbf{N}, \end{aligned} \tag{2}$$

is called a *linear congruential sequence* (for a thorough description, see [8]), which is a pseudo-random number generator. The output of Eq. (2) feeds into the function δ (in the second equation of the system) defined by

$$\delta(x) = \begin{cases} 0, & \text{if } x \neq d, \\ \\ 1, & \text{if } x = d, \end{cases} \tag{3}$$

where d is chosen from the set $\{0, 1, \ldots, m-1\}$. Under appropriate conditions on a, b, m, and with $x_0 \in \{0, 1, \ldots, m-1\}$, Eq. (2) generates a periodic sequence of whole numbers with period m that is some permuted version of the m-cycle $(0, 1, \ldots, m-1)$. The appropriate conditions are given by the following theorem (see Theorem A in Sect. 3.2.1.2 in [8]):

Theorem 1. *Let $\{x_n\}_{n=0}^{\infty}$ be a solution of Eq. (2) with $x_0 \in \mathbf{N} \cup \{0\}$. Then $\{x_n\}_{n=0}^{\infty}$ is periodic (or eventually periodic) with period m if and only if*

(i) the greatest common divisor gcd$(b, m) = 1$;
(ii) $a - 1$ is a multiple of every prime number p dividing m;
(iii) $a - 1$ is a multiple of 4 if m is a multiple of 4.

The second equation in system (1),

$$z_{n+1} = \begin{cases} \dfrac{-z_n - z_{n-1}}{2}, & \text{if } z_n + z_{n-1} \text{ is even}, \\ \\ -z_n - z_{n-1}, & \text{if } z_n + z_{n-1} \text{ is odd} \end{cases} + s\delta(x_n), \quad n = 0, 1, \ldots, \tag{4}$$
$$z_{-1}, z_0 \in \mathbf{Z}, \quad s \in \mathbf{Z} - \{0\},$$

is a modified version of one of the sixteen *Collatz difference equations* investigated by Amleh and colleagues (see Eq. (6*) in [1]):

$$y_{n+1} = \begin{cases} \dfrac{-y_n - y_{n-1}}{2}, & \text{if } y_n + y_{n-1} \text{ is even}, \\ \\ -y_n - y_{n-1}, & \text{if } y_n + y_{n-1} \text{ is odd}, \end{cases} \quad n = 0, 1, \ldots, \tag{5}$$
$$y_{-1}, y_0 \in \mathbf{Z}.$$

An important result stated in [1] associated with Eq. (5) is as follows:

Theorem 2. *Let $\{y_n\}_{n=-1}^{\infty}$ be a solution of Eq. (5) with $y_{-1}, y_0 \in \mathbf{Z}$. Suppose that the greatest common odd divisor gcod$(y_{-1}, y_0) = c$. Then $\{y_n\}_{n=-1}^{\infty}$ is eventually periodic in the three-cycle $(c, 0, -c)$ or the three-cycle $(-c, 0, c)$. In particular, if gcod$(y_{-1}, y_0) = 1$, then $\{y_n\}_{n=-1}^{\infty}$ is eventually periodic in the three cycle $(1, 0, -1)$ or the three cycle $(-1, 0, 1)$.*

Now, while it is observed (and conjectured) that every solution $\{z_n\}_{n=-1}^{\infty}$ of system (1) is eventually periodic, every solution is not eventually the three-cycle

$(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar. Instead, if m in Eq. (2) is sufficiently large, every solution $\{z_n\}_{n=-1}^{\infty}$ of system (1) contains intermittent or periodic recurrences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$). For, each time there is a developing three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$) in $\{z_n\}_{n=-1}^{\infty}$, the three-cycle is (in most cases) terminated by the addition of the nonzero value $s\delta(x_n) = s$ to $(-z_n - z_{n-1})/2$ or $-z_n - z_{n-1}$, where $\delta(x_n)$ is periodic with period m such that $\delta(x_n) = 1$ every m terms and $\delta(x_n) = 0$ otherwise. However, as will be seen in the sequel, for m in Eq. (2) sufficiently small, one does not even see interrupted three-cycles $(c, 0, -c)$ (or $(-c, 0, c)$).

So when m in Eq. (2) is sufficiently large, these periodic recurrences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar, in a solution $\{z_n\}_{n=-1}^{\infty}$ of system (1) are intended to represent the recurrent episodes of mood disturbance seen in an untreated individual with bipolar disorder. The addition of the nonzero value $s\delta(x_n) = s$ to $(-z_n - z_{n-1})/2$ or $-z_n - z_{n-1}$ is intended to represent either external environmental stimuli or internal biochemical stimuli which generally brings about temporary remission from an episode of mood disturbance.

We consider another modified version of the Collatz equation (which in this case is not part of a system) to serve as our second model of the longitudinal course of *treated* bipolar disorder:

$$u_{n+1} = \begin{cases} \dfrac{-u_n - u_{n-1}}{2}, & \text{if } u_n + u_{n-1} \text{ is even,} \\[2ex] -u_n - u_{n-1}, & \text{if } u_n + u_{n-1} \text{ is odd,} \end{cases} + t, \quad n = 0, 1, \ldots, \tag{6}$$

where $t \in \mathbf{Z} - \{0\}$ and $u_{-1}, u_0 \in \mathbf{Z}$. We are able to show that every solution $\{u_n\}_{n=-1}^{\infty}$ of Eq. (6) is eventually periodic, but not eventually the three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar, and contains no occurences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$). Therefore, a solution $\{u_n\}_{n=-1}^{\infty}$ is intended to represent the period of remission that an individual with bipolar disorder is in while undergoing treatment with medication, in particular, with what is referred to as a *mood-stabilizing drug* possibly along with an antidepressant and some other psychiatric drugs.

In Sect. 2, we consider three examples along with their associated results or conjectures. The first example follows the longitudinal progression of bipolar disorder in an untreated individual. System (1) is assigned values for its parameters and initial conditions, and m is sufficiently large enough to generate an eventually periodic solution $\{z_n\}_{n=-1}^{\infty}$ with periodic recurrences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar. We subsequently show that if an arbitrary solution $\{z_n\}_{n=-1}^{\infty}$ of system (1) is eventually periodic with period p, then $m|p$. We conjecture that every solution of system (1) is eventually periodic (but not necessarily with the same period).

The second example is that of a disease-free first-degree relative (i.e., a relative who is a parent, a full sibling, or an offspring) of an individual with bipolar disorder. Since bipolar disorder is a genetic disease and thus hereditary, a first-degree relative is most likely predisposed to but does not have to be stricken with the disease bipolar disorder. System (1) is assigned values for its parameters and initial conditions, and m is sufficiently small enough to generate an eventually periodic solution $\{z_n\}_{n=-1}^{\infty}$ with no periodic occurrences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar.

The third example follows the longitudinal course of an individual with bipolar disorder who is chronically treated with medication. System (6) is used and its parameter and initial conditions are assigned values. There are absolutely no occurrences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar, in the eventually periodic solution $\{u_n\}_{n=-1}^{\infty}$. We then show that every solution $\{u_n\}_{n=-1}^{\infty}$ of Eq. (6) is bounded (with the bound dependent on initial values). We thus have bounded sequences of integers generated by a difference equation, and so every solution of Eq. (6) is eventually periodic (but not necessarily with the same period).

Section 3 offers a neuropsychiatric interpretation of the examples and results in Sect. 2.

2 Examples, Results, and Conjectures

2.1 Untreated Bipolar Disorder

Here we first give an example of a solution $\{z_n\}_{n=-1}^{\infty}$ of our first model, system (1), which represents the longitudinal course of the disease, bipolar disorder, of an untreated individual.

Example 1. We assign values to the parameters and initial conditions of the linear congruential sequence, Eq. (2), and modified Collatz equation, Eq. (4), in system (1), together with prescribing d in the definition of δ in Eq. (3), as follows:

1. We let $a = 21$, $b = 3$, $m = 20$, and $x_0 = 1$ in Eq. (2), which satisfies Theorem 1, thereby making the output, $\{x_n\}_{n=0}^{\infty}$, periodic with period 20. Then the first equation in system (1) is

$$x_{n+1} = (21x_n + 3) \bmod 20, \quad n = 0, 1, \ldots.$$

2. We let $z_{-1} = 7$, $z_0 = 10$, and $s = 20$ (the fact that $m = 20$ and $s = 20$ is just coincidence) in Eq. (4). Then the second equation in system (1) is

$$z_{n+1} = \begin{cases} \dfrac{-z_n - z_{n-1}}{2}, & \text{if } z_n + z_{n-1} \text{ is even,} \\ -z_n - z_{n-1}, & \text{if } z_n + z_{n-1} \text{ is odd} \end{cases} + 20\delta(x_n), \quad n = 0, 1, \ldots.$$

3. We let $d = 15 \in \{0, 1, \ldots, m-1 = 19\}$ in Eq. (3) to give us

$$\delta(x) = \begin{cases} 0, & \text{if } x \neq 15, \\ 1, & \text{if } x = 15. \end{cases}$$

   Note that then the solution $\{x_n\}_{n=0}^{\infty}$ of the linear congruential sequence is such that for $k = 0, 1, \ldots, x_{20k+18} = 15$ so that $\delta(x_{20k+18}) = 1$ (and $\delta(x_n) = 0$ otherwise), and so, in turn, the terms z_{20k+19} in the solution of the modified Collatz equation have 20 added to $(-z_n - z_{n-1})/2$ or $-z_n - z_{n-1}$.

The first 83 terms of $\{z_n\}_{n=-1}^{\infty}$ are as follows:

$$\begin{aligned}
&z_{-1} = 7,\ z_0 = 10,\ z_1 = -17,\ z_2 = 7,\ z_3 = 5,\ z_4 = -6,\ z_5 = 1,\\
&z_6 = 5,\ z_7 = -3,\ z_8 = -1,\ z_9 = 2,\ z_{10} = -1,\ z_{11} = -1,\ z_{12} = 1,\\
&z_{13} = 0,\ z_{14} = -1,\ z_{15} = 1,\ z_{16} = 0, \parallel z_{17} = -1,\ z_{18} = 1,\ z_{19} = 0 + 20 = 20,\\
&z_{20} = -21,\ z_{21} = 1,\ z_{22} = 10,\ z_{23} = -11,\ z_{24} = 1,\ z_{25} = 5,\ z_{26} = -3,\\
&z_{27} = -1,\ z_{28} = 2,\ z_{29} = -1,\ z_{30} = -1,\ z_{31} = 1,\ z_{32} = 0,\ z_{33} = -1,\\
&z_{34} = 1,\ z_{35} = 0,\ z_{36} = -1,\ z_{37} = 1,\ z_{38} = 0,\ z_{39} = -1 + 20 = 19,\ z_{40} = -19,\\
&z_{41} = 0,\ z_{42} = 19,\ z_{43} = -19,\ z_{44} = 0,\ z_{45} = 19,\ z_{46} = -19,\ z_{47} = 0,\\
&z_{48} = 19,\ z_{49} = -19,\ z_{50} = 0,\ z_{51} = 19,\ z_{52} = -19,\ z_{53} = 0,\ z_{54} = 19,\\
&z_{55} = -19,\ z_{56} = 0,\ z_{57} = 19,\ z_{58} = -19,\ z_{59} = 0 + 20 = 20,\ z_{60} = -1,\ z_{61} = -19,\\
&z_{62} = 10,\ z_{63} = 9,\ z_{64} = -19,\ z_{65} = 5,\ z_{66} = 7,\ z_{67} = -6,\ z_{68} = -1,\\
&z_{69} = 7,\ z_{70} = -3,\ z_{71} = -2,\ z_{72} = 5,\ z_{73} = -3,\ z_{74} = -1,\ z_{75} = 2,\\
&z_{76} = -1, \parallel z_{77} = -1,\ z_{78} = 1,\ z_{79} = 0 + 20 = 20,\ z_{80} = -21,\ z_{81} = 1,\ \ldots.
\end{aligned}$$

We describe some prominent features of the solution $\{z_n\}_{n=-1}^{\infty}$ and present some interpretations:

1. Since $z_{77} = z_{17}, z_{78} = z_{18}$, and $20\delta(x_{78}) = 20\delta(x_{18}) = 20$, we have that $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic with period 60, i.e., $\{z_n\}_{n=17}^{\infty}$ is periodic with period 60. Note that $m|60$, where $m = 20$.
2. Since $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic, we can say that the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar, recurs indefinitely or periodically. Specifically, the interrupted three-cycle $(1, 0, -1)$ or $(-19, 0, 19)$ occurs with the blocks of successive terms $z_{12} - z_{18}$, $z_{60k+31} - z_{60k+38}$, for $k = 0, 1, \ldots$, and $z_{60k+40} - z_{60k+58}$, for $k = 0, 1, \ldots$. These intermittent or periodic recurrences of the interrupted three-cycle $(1, 0, -1)$ or $(-19, 0, 19)$ are intended to represent the recurrent episodes of mood disturbance in the untreated bipolar individual, and all other terms in $\{z_n\}_{n=-1}^{\infty}$ are intended to represent remission of the disease.
3. In $\{z_n\}_{n=-1}^{\infty}$, immediately prior to each occurrence of the interrupted three-cycle $(1, 0, -1)$, is the block of five consecutive terms $-3, -1, 2, -1, -1$. Specifically, this block of five consecutive terms occurs with $z_7 - z_{11}$ and $z_{60k+26} - z_{60k+30}$, for $k = 0, 1, \ldots$. This block of terms, $3, -1, 2, -1, -1$, is intended to represent the occurrence of stressful events which precipitate most episodes of mood disturbance in the untreated bipolar individual.

4. The terms of $\{z_n\}_{n=-1}^{\infty}$ in which there is the periodic addition of $20\delta(x_n) = 20$ are z_{19}, z_{60k+39}, for $k = 0, 1, \ldots$, and z_{60k+59}, for $k = 0, 1, \ldots$. In particular, the terms z_{19} and z_{60k+59}, for $k = 0, 1, \ldots$, are intended to represent the input of external environmental or internal biochemical stimuli, which in some way terminate the episodes of mood disturbance in the untreated bipolar individual. The terms z_{60k+39}, for $k = 0, 1, \ldots$, are intended to represent external or internal stimuli which in some way exacerbate episodes of mood disturbance. □

We next state a result and conjecture, with the parameters and initial conditions kept arbitrary.

Theorem 3. *Let $\{z_n\}_{n=-1}^{\infty}$ be a solution of system (1), with $\{x_n\}_{n=0}^{\infty}$ a solution of the linear congruential sequence of system (1) which satisfies Theorem 1. If $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic with period q, then q is a multiple of m, the period of $\{x_n\}_{n=0}^{\infty}$ and thus of $\{s\delta(x_n)\}_{n=0}^{\infty}$ in the modified Collatz equation of system (1).*

Proof. For the sake of convenience, we use the fact that if q is a multiple of m, then the least common multiple $\operatorname{lcm}(m, q) = q$, which, in turn, implies that if q is not a multiple of m, then $\operatorname{lcm}(m, q) = kq$, for $k \in \{2, 3, \ldots\}$.

Now by hypothesis, we have that $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic with period q. So, for the sake of contradiction, we suppose that q is not a multiple of m and $\operatorname{lcm}(m, q) = kq$, for $k \in \{2, 3, \ldots\}$. Then there exists $N \geq 1$ such that

1. $\{z_n\}_{n=N}^{\infty}$ is periodic with period q.
2. $\delta(x_{N-1}) = 1$.

Consequently, since $\operatorname{lcm}(m, q) \neq q$ but $\operatorname{lcm}(m, q) = kq$, for $k \in \{2, 3, \ldots\}$, $\delta(x_{N+q-1}) = 0$ but $\delta(x_{N+kq-1}) = 1$. Then, for some $\alpha, \beta \in \mathbf{Z}$, $z_{N+q-2} = \alpha$, $z_{N+q-1} = \beta$, and

$$z_{N+q} = \begin{cases} \dfrac{-\alpha - \beta}{2}, & \text{if } \alpha + \beta \text{ is even,} \\ \\ -\alpha - \beta, & \text{if } \alpha + \beta \text{ is odd} \end{cases} + 0 \equiv \gamma.$$

By periodicity of $\{z_n\}_{n=N}^{\infty}$ with period q, we have that

$$z_{N+kq-2} = \alpha, \quad z_{N+kq-1} = \beta, \quad z_{N+kq} = \gamma.$$

On the other hand, since $\delta(x_{N+kq-1}) = 1$, then

$$z_{N+kq} = \gamma + s, \quad s \neq 0.$$

Therefore,

$$\gamma = z_{N+kq} = \gamma + s,$$

which gives us a contradiction. Hence, q is a multiple of m. □

Conjecture 1. Let $\{z_n\}_{n=-1}^{\infty}$ be a solution of system (1), and suppose that the linear congruential sequence in system (1) satisfies Theorem 1. Then $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic.

2.2 First-Degree Relative of Someone with Bipolar Disorder

We give an example of a solution $\{z_n\}_{n=-1}^{\infty}$ of our first model, system (1), which is similar to Example 1, except that m in the linear congruential sequence is smaller than $m = 20$ (and so the parameters a, b are different too). We intend to have this second example represent a first-degree relative of someone who has bipolar disorder. We assume that the relative does not have the disease, but has the genotype (i.e., genetic makeup) predisposing one to have the disease. We follow the longitudinal course of the disease-free life of this first-degree relative.

Example 2. We assign values to the parameters and initial conditions of the linear congruential sequence, Eq. (2), and modified Collatz equation, Eq. (4), in system (1), along with prescribing d in the definition of δ in Eq. (3), as follows:

1. We let $a = 11$, $b = 3$, $m = 10$, and $x_0 = 1$ in Eq. (2), which satisfies Theorem 1, thereby making the output, $\{x_n\}_{n=0}^{\infty}$, periodic with period 10. Then the first equation in system (1) is

$$x_{n+1} = (11x_n + 3) \bmod 10, \quad n = 0, 1, \ldots.$$

2. We let $z_{-1} = 7$, $z_0 = 10$, and $s = 20$ in Eq. (4). Then the second equation in system (1) is

$$z_{n+1} = \begin{cases} \dfrac{-z_n - z_{n-1}}{2}, & \text{if } z_n + z_{n-1} \text{ is even,} \\ -z_n - z_{n-1}, & \text{if } z_n + z_{n-1} \text{ is odd} \end{cases} + 20\delta(x_n), \quad n = 0, 1, \ldots.$$

3. We let $d = 5 \in \{0, 1, \ldots, m-1 = 9\}$ in Eq. (3) to give us

$$\delta(x) = \begin{cases} 0, & \text{if } x \neq 5, \\ 1, & \text{if } x = 5. \end{cases}$$

Note that then the solution $\{x_n\}_{n=0}^{\infty}$ of the linear congruential sequence is such that for $k = 0, 1, \ldots$, $x_{10k+8} = 5$ so that $\delta(x_{10k+8}) = 1$ (and $\delta(x_n) = 0$ otherwise), and so, in turn, the terms z_{10k+9} in the solution of the modified Collatz equation have 20 added to $(-z_n - z_{n-1})/2$ or $-z_n - z_{n-1}$.

The first 53 terms of $\{z_n\}_{n=-1}^{\infty}$ are as follows:

$$\begin{gathered}z_{-1}=7,\ z_0=10,\ z_1=-17,\ z_2=7,\ z_3=5,\ z_4=-6,\\ z_5=1,\ z_6=5,\ z_7=-3,\ z_8=-1,\ z_9=2+20=22,\ z_{10}=-21,\\ z_{11}=-1,\ z_{12}=11,\ z_{13}=-5,\ z_{14}=-3,\ z_{15}=4,\ z_{16}=-1,\\ z_{17}=-3,\ z_{18}=2,\ z_{19}=1+20=21,\ z_{20}=-23,\ z_{21}=1,\ z_{22}=11,\\ z_{23}=-6,\ z_{24}=-5,\ z_{25}=11,\ z_{26}=-3,\ z_{27}=-4,\ \|\ z_{28}=7,\\ z_{29}=-3+20=17,\ z_{30}=-12,\ z_{31}=-5,\ z_{32}=17,\ z_{33}=-6,\ z_{34}=-11,\\ z_{35}=17,\ z_{36}=-3,\ z_{37}=-7,\ z_{38}=5,\ z_{39}=1+20=21,\ z_{40}=-13,\\ z_{41}=-4,\ z_{42}=17,\ z_{43}=-13,\ z_{44}=-2,\ z_{45}=15,\ z_{46}=-13,\\ z_{47}=-1,\ \|\ z_{48}=7,\ z_{49}=-3+20=17,\ z_{50}=-12,\ z_{51}=-5,\ \ldots.\end{gathered}$$

We describe two prominent features of the solution $\{z_n\}_{n=-1}^{\infty}$, along with presenting our interpretations:

1. Since $z_{48}=z_{28}, z_{49}=z_{29}$, and $20\delta(x_{48})=20\delta(x_{28})=20$, we have that $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic with period 20, i.e., $\{z_n\}_{n=28}^{\infty}$ is periodic with period 20. Note that $m|20$, where $m=10$.
2. Since $\{z_n\}_{n=-1}^{\infty}$ is eventually periodic, we can say with definitude that there are no occurrences of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar. This lack of occurrence of the interrupted three-cycle is intended to represent the disease-free state that the first-degree relative is in. This completely healthy state is evidently due to the increased frequency of input of external environmental or internal biochemical stimuli, which abolishes the occurrence of episodes of mood disturbance and which is represented by the increased frequency of addition of $20\delta(x_n)=20$ because m is relatively small (m is 10 as against 20). □

2.3 Treated Bipolar Disorder

We first give an example of a solution $\{u_n\}_{n=-1}^{\infty}$ of our second model, Eq. (6), which represents the longitudinal course of the disease, bipolar disorder, of an individual treated with mood-stabilizing drugs.

Example 3. We let $u_{-1}=7, u_0=10$, and $t=4$ in this second model to give us the equation

$$u_{n+1}=\begin{cases}\dfrac{-u_n-u_{n-1}}{2}, & \text{if } u_n+u_{n-1} \text{ is even,}\\[2ex] -u_n-u_{n-1}, & \text{if } u_n+u_{n-1} \text{ is odd}\end{cases} +4, \quad n=0,1,\ldots.$$

The solution $\{u_n\}_{n=-1}^{\infty}$ is then the following:

$u_{-1} = 7,\ u_0 = 10,\ u_1 = -17 + 4 = -13,\ u_2 = 3 + 4 = 7,$
$u_3 = 3 + 4 = 7,\ u_4 = -7 + 4 = -3,\ u_5 = -2 + 4 = 2,\ u_6 = 1 + 4 = 5, \|$
$u_7 = -7 + 4 = -3,\ u_8 = -1 + 4 = 3,\ u_9 = 0 + 4 = 4, \|\ u_{10} = -7 + 4 = -3,$
$u_{11} = -1 + 4 = 3,\ u_{12} = 0 + 4 = 4,\ \ldots.$

We make the following observations on and interpretation of the solution $\{u_n\}_{n=-1}^{\infty}$:

1. Clearly, since $u_{10} = u_7$ and $u_{11} = u_8$, we have that $\{u_n\}_{n=-1}^{\infty}$ is eventually periodic with period 3, i.e., $\{z_n\}_{n=7}^{\infty}$ is periodic with period 3.
2. Since $\{u_n\}_{n=-1}^{\infty}$ is eventually periodic, we can say with certainty that there is the complete absence of the interrupted three-cycle $(c, 0, -c)$ (or $(-c, 0, c)$), c an odd scalar. This absence is intended to represent the successful treatment by medications (i.e., the complete abolishment of episodes of mood disturbance by drugs) of the individual with bipolar disorder. Note that the addition of 4 at every iteration in the computation of terms of the solution $\{u_n\}_{n=-1}^{\infty}$ is intended to represent the continuous and constant medication regimen that the compliant bipolar individual is on, which in an ideal world prevents any recurrence of episodes of mood disturbance. □

We next give results:

Remark 1. Let $\{u_n\}_{n=-1}^{\infty}$ be a solution of Eq. (6), and let c be an odd scalar. Then there is no occurrence of the block of three successive terms $[c, 0, -c]$ or $[-c, 0, c]$ in the sequence $\{u_n\}_{n=-1}^{\infty}$.

We give the proof for the absence of the block $[c, 0, -c]$. The proof for the absence of the block $[-c, 0, c]$ is similar and will be omitted.

Proof. If we assume the contrary (i.e. there is at least one occurrence of the block of three successive terms $[c, 0, -c]$), we can further assume that there exists $N \geq -1$ such that

$$u_N = c, \quad u_{N+1} = 0, \quad u_{N+2} = -c.$$

Then, on the other hand, we can compute u_{N+2} from u_N, which we know is odd, and u_{N+1} which we know is even:

$$u_{N+2} = -u_{N+1} - u_N + t = -c + t, \quad t \neq 0.$$

Therefore,

$$-c = u_{N+2} = -c + t,$$

which gives us a contradiction. □

We now show that every solution $\{u_n\}_{n=-1}^{\infty}$ of Eq. (6) is eventually periodic; but to do so, we first need to show that every solution is bounded.

Lemma 1. *Let* $\{u_n\}_{n=-1}^{\infty}$ *be a solution of Eq. (6). Then*

$$|u_n| \leq |u_{-1}| + |u_0| + |t|, \quad \textit{for all } n \geq -1.$$

Proof. We will prove by induction that for every $n \geq 0$,

$$|u_{n-1}|, \quad |u_n|, \quad \left|\frac{-u_n - u_{n-1}}{2} + t\right|,$$

$$|-u_n - u_{n-1} + t| \in [0, |u_{-1}| + |u_0| + |t|].$$

The claim is clearly true for $n = 0$.

So, suppose that $n \geq 0$ and that

$$|u_{n-1}|, \quad |u_n|, \quad \left|\frac{-u_n - u_{n-1}}{2} + t\right|,$$

$$|-u_n - u_{n-1} + t| \in [0, |u_{-1}| + |u_0| + |t|].$$

We will show that

$$|u_n|, \quad |u_{n+1}|, \quad \left|\frac{-u_{n+1} - u_n}{2} + t\right|,$$

$$|-u_{n+1} - u_n + t| \in [0, |u_{-1}| + |u_0| + |t|].$$

By the inductive hypothesis,

$$|u_n| \leq |u_{-1}| + |u_0| + |t|.$$

Also by the inductive hypothesis,

$$|u_{n+1}| = \left|\frac{-u_n - u_{n-1}}{2} + t\right| \leq |u_{-1}| + |u_0| + |t|,$$

if $u_n + u_{n-1}$ is even; and

$$|u_{n+1}| = |-u_n - u_{n-1} + t| \leq |u_{-1}| + |u_0| + |t|,$$

if $u_n + u_{n-1}$ is odd.

We next need to show that $\left|\frac{-u_{n+1} - u_n}{2} + t\right| \in [0, |u_{-1}| + |u_0| + |t|]$:

If $u_n + u_{n-1}$ is even, then

$$\begin{aligned}\left|\frac{-u_{n+1}-u_n}{2}+t\right| &= \left|\frac{-\left(\dfrac{-u_n-u_{n-1}}{2}+t\right)-u_n}{2}+t\right| \\ &\le \tfrac{1}{4}|u_n|+\tfrac{1}{4}|u_{n-1}|+\tfrac{1}{2}|t| \\ &\le |u_{-1}|+|u_0|+|t|.\end{aligned}$$

If u_n+u_{n-1} is odd, then

$$\begin{aligned}\left|\frac{-u_{n+1}-u_n}{2}+t\right| &= \left|\frac{-(-u_n-u_{n-1}+t)-u_n}{2}+t\right| \\ &\le \tfrac{1}{2}|u_{n-1}|+\tfrac{1}{2}|t| \\ &\le |u_{-1}|+|u_0|+|t|.\end{aligned}$$

Finally, we show that $|-u_{n+1}-u_n+t| \in [0, |u_{-1}|+|u_0|+|t|]$:
If u_n+u_{n-1} is even, then

$$\begin{aligned}|-u_{n+1}-u_n+t| &= \left|-\left(\frac{-u_n-u_{n-1}}{2}+t\right)-u_n+t\right| \\ &\le \tfrac{1}{2}|u_n|+\tfrac{1}{2}|u_{n-1}| \\ &\le |u_{-1}|+|u_0|+|t|.\end{aligned}$$

If u_n+u_{n-1} is odd, then

$$\begin{aligned}|-u_{n+1}-u_n+t| &= |-(-u_n-u_{n-1}+t)-u_n+t| \\ &\le |u_{n-1}| \\ &\le |u_{-1}|+|u_0|+|t|.\end{aligned}$$

□

Theorem 4. *Let $\{u_n\}_{n=-1}^{\infty}$ be a solution of Eq. (6). Then $\{u_n\}_{n=-1}^{\infty}$ is eventually periodic.*

Proof. By Lemma 1, there exists $M>0$ such that $|u_n|\le M$ for all $n\ge -1$. Thus, $\{u_n\}_{n=-1}^{\infty}$ is a bounded sequence of integers, which are generated by a second-order difference equation. It then follows that there exist $N\ge -1$ and $q\in\{2,3,\ldots\}$ such that

$$u_{N+q}=u_N \quad \text{and} \quad u_{N+1+q}=u_{N+1}.$$

We then have that $\{u_n\}_{n=N}^{\infty}$ is periodic with period q. □

3 Concluding Remarks

The literature is rife with speculations and investigations into the etiology of bipolar disorder. We find that two particular proposals on what underlies bipolar disorder stand out. One proposal is that bipolar disorder is the result of inherited abnormalities in the cellular signaling networks of the brain (i.e., the biochemical pathways and cascades by which there is interneuronal communication or communication between neurons). These abnormalities, in turn, lead to widespread dysfunction in a variety of physiological processes which comprise bipolar disorder. (See [5] for the details on the cellular signaling cascades felt to be involved, as well as [2, 4, 6, 9, 10], and [15] for discussions on the neurochemical and physiological processes affected by the cellular signaling dysfunction.)

The other proposal is that in bipolar individuals there are abnormalities in sleep and in the endogenous genetic circadian clock, which cause endocrinologic, biochemical, and electrophysiological disturbances, which, in turn, manifest themselves as the signs and symptoms of bipolar disorder. (For a general discussion, see [5], and for examples, see [7, 11–14], and [16].)

Based on our examples and results in Sect. 2, we now speculate on one aspect of what causes the episodes of mood disturbance seen in bipolar disorder, at least early on in the course of the disease. We feel this aspect, environmental in nature, interacts with the genetic predisposition to having bipolar disorder.

We offer the proposition, which is actually the revisiting of an old concept, that the episodes of mood disturbance that define bipolar disorder are triggered by either overt or obscure precipitating environmental events, peculiar in content to the individual undergoing the episodes, at least early on in the disease. Therefore, we believe that there is indeed a delicate interplay between a stressful environment and the dysfunctional biochemistry/physiology of an individual with bipolar disorder. We refer to the precipitating event as a *traumatic event*.

We define a traumatic event to be an environmental event characterized by external stimuli that cannot be fully processed by the brain, i.e., external stimuli generating internal informational thoughts which are then transmitted along neurons that are not, for a particular individual, part of any recursive network of neurons. We then contend that this inability to process external stimuli internally leaves a void in mental activity that then triggers compensatory biochemical and physiological processes, only in the case of a bipolar individual the processes, which were inherited, are dysfunctional.

We then propose that either treatment of the bipolar individual with mood-stabilizing drugs or reception by the bipolar individual of incoming new informational stimuli that make it possible for the brain to assimilate the previously received traumatic event, can bring about remission from the current episode of mood disturbance.

References

1. Amleh, A.M., Grove, E.A., Kent, C.M., Ladas, G.: On some difference equations with eventually periodic solutions. J. Math. Anal. **223**, 196–215 (1998)
2. Atagun, M.I., Sikoglu, E.M., Can, S.S., Ugurla, G.K., Kaymak, S.U., Caykoylu, A., Algin, O., Phillips, M.L., Moore, C.M., Ongur, D.: Neurochemical differences between bipolar disorder type I and II in superior temporal cortices: a proton magnetic resonance spectroscopy study. J. Affect. Disord. **235**, 15–19 (2018)
3. Frances, A. (Chairperson of Task Force): Diagnostic and Statistical Manual of Mental Disorders, and text revision. 4th edn. American Psychiatric Association, Washington, DC (2000)
4. Galinska-Skok, B., Konarzewska, B., Kubas, B., Tarasow, E., Szule, A.: Neurochemical alterations in anterior cingulate cortex in bipolar disorder: a proton magnetic resonance spectroscopy study (1H-MRS). Psychiatr. Pol. **50**(4), 839–848 (2016)
5. Goodwin, F.K., Jamison, K.R.: Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression. Oxford University Press, New York (2007)
6. Hajek, T., Bauer, M., Pfenning, A., Cullis, J., Ploch, J., O'Donovan, C., Bohner, G., Klingebiel, R., Young, L.T., MacQueen, G.M., Alda, M.: Large positive effect of lithium on prefrontal cortex N-acetylaspartate in patients with bipolar disorder: 2-centre study. J. Psychiatry Neurosci. **37**(3), 185–192 (2012)
7. Harvey, A.G., Talbot, L.S., Gershon, A.: Sleep disturbance in bipolar disorder across the lifespan. Clin. Psychol. (New York) **16**(2), 256–277 (2009)
8. Knuth, D.E.: The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3rd edn. Addison-Wesley, Boston (1998)
9. Kulak, A., Steullet, P., Cabungcal, J., Werge, T., Ingason, A., Cuenod, M., Quang Do, K.: Redox dysregulation in the pathophysiology of schizophrenia and bipolar disorder: insights from animal models. Antioxid. Redox Signal. **18**(12), 1428–1443 (2013)
10. Manji, H.K., Quiroz, J.A., Payne, J.L., Singh, J., Lopes, B.P., Viegas, J.S., Zarate, C.A.: The underlying neurobiology of bipolar disorder. World Psychiatry **2**(3), 136–146 (2003)
11. Mansour, H.A., Monk, T.H., Nimgaonnkar, V.L.: Circadian genes and bipolar disorder. Ann. Med. **37**, 196–205 (2005)
12. Moon, J., Cho, C., Son, G.H., Geum, D., Chung, S., Kim, H., Kang, S., Park, Y., Yoon, H., Kim, L., Jee, H., An, H., Kripke, D.F., Lee, H.: Advanced circadian phase in mania and delayed circadian phase in mixed mania and depression returned to normal after treatment of bipolar disorder. EBioMedicine **11**, 285–295 (2016)
13. Plante, D.T., Winkelman, J.W.: Sleep disturbance in bipolar disorder: therapeutic implications. Am. J. Psychiatry **165**, 830–843 (2008)
14. Rumble, M.E., White, K.H., Benca, R.M.: Sleep disturbances in mood disorders. Psychiatr. Clin. N. Am. **38**, 743–759 (2015)
15. Stall, A.L., Sachs, G.S., Cohen, B.M., Lafer, B., Christensen, Renshaw, P.F.: Choline in the treatment of rapid-cycling bipolar disorder: clinical and neurochemical findings in lithium-treated patients. Biol. Psychiatry **40**, 382–388 (1996)
16. Takaesu, Y., Inoue, Y., Murakoshi, A., Komada, Y., Otsuka, A., Fulenma, K., Inoue, T.: Prevalence of circadian rhythm sleep-wake disorders and associated factors in euthymic patients with bipolar disorder. PLoS ONE **11**(7), e0159578 (2016). https://doi.org/10.1371/journal.pone.0159578

Local Bifurcations in the Generalized Cahn-Hilliard Equation

A. Kulikov and D. Kulikov

Abstract A periodic boundary value problem for a generalized Cahn-Hilliard equation is studied. Bifurcation problems are considered. The analysis of these bifurcation problems use the methods of invariant manifold and the Poincare normal forms for the dynamic systems with an infinite-dimensional space of initial conditions. It is proved that this dynamic systems has a local attractor formed by unstable solutions in the sense of Lyapunov definition. Asymptotic formulas for these solutions are obtained.

Keywords Bifurcations · Generalized Cahn-Hilliard equation · Periodic boundary value problem

1 Introduction

The paper considers one of the well-known equations of mathematical physics

$$u_t - c(u^2)_x + (u_{xx} + bu + b_2u^2 - b_3u^3)_{xx} = 0, \tag{1}$$

where $u = u(t, x)$ and a, b, b_2, b_3, c are real constants, $b_3 \geq 0$.

This partial differential equation (PDE) is commonly called the generalized Cahn-Hilliard equation. This nonlinear equation is used to describe various phenomena in physics, hydrodynamics and chemical kinetics.

For $c = b_2 = 0$, we obtain the original version of such an equation, which was proposed in [1] and describes the evolution of the interface between two substances in the case when a chemical reaction takes place between them.

If $c \neq 0$, then Eq. (1) is usually called the convective version of the Cahn-Hilliard equation [2].

A. Kulikov · D. Kulikov (✉)
Demidov Yaroslavl State University, Yaroslavl, Russia
e-mail: kulikov_d_a@mail.ru

A. Kulikov
e-mail: anat_kulikov@mail.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_14

If $c = b_3 = 0$, then we obtain a version of the Cahn-Hilliard equation, which sometimes is called the Pukhnachev equation. This version arose in hydrodynamics [3, 4], and it describes the change in the interface between two liquids.

Following many works (see, for example, [1–6]), we supplement Eq. (1) with periodic boundary conditions. Without loss of generality, we can assume that these conditions have the form

$$u(t, x + 2\pi) = u(t, x). \tag{2}$$

We note that the boundary value problem (1), (2) has a family of homogeneous equilibrium states $u(t, x) = const$. Next, we consider the existence and stability of solutions of the boundary value problem (1), (2), which essentially depend on x ($u_x(t, x) \neq 0$) and are close to homogeneous equilibrium states.

Remark 1. The Cahn-Hilliard equation maybe studied with other boundary conditions (see, for example, [7]).

If we supplement the boundary value problem (1), (2) with the initial condition

$$u(0, x) = f(x), \tag{3}$$

then we obtain the mixed problem (1), (2), (3) which is locally correctly solvable, if $f(x) \in H_4$ [8]. Recall that H_4 denotes the Sobolev functional space [9], containing 2π- periodic function $f(x)$, which have generalized derivatives up to fourth order $f'(x), f''(x), f'''(x), f^{(IV)} \in L_2(0, 2\pi)$. Recall that in this situation it follows from the embedding theorems that $f(x) \in C^3[0, 2\pi]$. Let

$$M_0(f) = \frac{1}{2\pi}\int\limits_0^{2\pi} f(x)dx,\ \ M_0(u(t, x)) = \frac{1}{2\pi}\int\limits_0^{2\pi} u(t, x)dx$$

be the spatial average.

Lemma 1. *The following identity holds*

$$M_0(u(t, x)) = \alpha,$$

where $\alpha \in \mathbb{R}$, $u(t, x)$ is a solution of the boundary value problem (1), (2).

The proof of the lemma is based on the integration of Eq. (1) taking into account the boundary conditions (2). It is clear that

$$\frac{1}{2\pi}\int\limits_0^{2\pi}(u^2)_x dx = 0,\ \frac{1}{2\pi}\int\limits_0^{2\pi}(u_{xx} + bu + b_2u^2 - b_3u^3)_{xx} dx = 0.$$

So,

$$\frac{1}{2\pi}\int_0^{2\pi} u_t(t,x)dx = \frac{d}{dt}(\frac{1}{2\pi}\int_0^{2\pi} u(t,x)dx) = 0.$$

Consequently, $\frac{1}{2\pi}\int_0^{2\pi} u(t,x)dx = \alpha \in \mathbb{R}$.

By $H_4(\alpha)$ we denote the affine space of functions $f(x)$, for which $M_0(f) = \alpha$. Obviously, $H_4(\alpha)$ is invariant for solutions of the boundary value problem (1), (2) in the following sense: if $f(x) \in H_4(\alpha)$, then the solution of the initial boundary value problem (1), (2), (3) $u(t,x) \in H_4(\alpha)$, for all t, when it exists. Of course, $u(t,x) \equiv \alpha$ also belongs to $H_4(\alpha)$.

Now, set

$$u(t,x) = \alpha + v(t,x). \tag{4}$$

Substitution (4) reduces the boundary value problem (1), (2) to a similar boundary value problem for the auxiliary function $v(t,x)$. Thus, we obtain

$$v_t + v_{xxxx} + b(\alpha)v_{xx} - c(\alpha)v_x - c(v^2)_x + [b_2(\alpha)v^2 - b_3v^3]_{xx} = 0, \tag{5}$$

$$v(t, x + 2\pi) = v(t,x),\ M_0(v) = 0. \tag{6}$$

Here, $b(\alpha) = b + 2\alpha b_2 - 3\alpha^2 b_3$, $b_2(\alpha) = b_2 - 3\alpha b_3$, $c(\alpha) = 2c\alpha$. The boundary value problem (5), (6) has a unique spatially homogeneous steady state $v(t,x) \equiv 0$. The main part of the work will be devoted to studying the behavior of solutions of the boundary value problem (5), (6) with initial conditions from a neighborhood of the zero solution. The neighborhood is understood in the sense of the norm of the phase space (the space of initial conditions), i.e. $H_{4,0}$, where $f(x) \in H_{4,0}$ if $f(x) \in H_4$ and $M_0(f) = 0$. In conclusion of this section, we emphasize that boundary value problems (1), (2) and (5), (6) can be included in the class of abstract parabolic equations (see, for example, [8, 10]).

2 Stability Analysis of the Zero Equilibrium State of the Auxiliary Boundary Value Problem

In this section, we consider a linearised version of the boundary value problem (5), (6), i.e. the following

$$v_t = A(\alpha)v, \tag{7}$$

$$v(t, x + 2\pi) = v(t,x),\ M_0(v) = 0, \tag{8}$$

where the linear differential operator (LDO) on the right side of Eq. (7) is defined by the equality

$$A(\alpha)y = -y^{(IV)} - b(\alpha)y'' + c(\alpha)y',\ y = y(x),\ y(x+2\pi) = y(x),\ M_0(y) = 0.$$

We shall be considering this LDO in the space $H_{0,0}$ which consists of 2π-periodic functions $y(x) \in L_2(0, 2\pi)$, if $x \in (0, 2\pi)$, and has zero spatial average, i.e. $M_0(y) = 0$. Regarding its domain, we can choose, for example, the set of 2π-periodic functions with zero spatial average. Obviously, in our case, LDO $A(\alpha)$ has countable set of eigenvalues

$$\lambda_n = \lambda_n(\alpha) = -n^4 + b(\alpha)n^2 + ic(\alpha)n,$$

with corresponding eigenfunctions $\exp(inx), n = \pm 1, \pm 2, \pm 3, \dots$ The family of functions $\{\exp(\pm inx)\}$ forms a complete orthogonal system in the space $L_{2,0}(0, 2\pi)$ (a function $f(x) \in L_{2,0}(0, 2\pi)$, if $f(x) \in L_2(0, 2\pi)$ and $M_0(f) = 0$). Therefore, the following statement holds.

Lemma 2. *The solutions of the boundary value problem (7), (8) are asymptotically stable if* $b(\alpha) < 1$ *and unstable, if* $b(\alpha) > 1$. *For* $b(\alpha) = 1$ *they are stable.*

Moreover, for $b(\alpha) < 1$ the zero solution to the nonlinear boundary value problem (5), (6) is asymptotically stable, and for $b(\alpha) > 1$. For $b(\alpha) = 1$ the critical case in the stability problem of the zero solution takes place.

The inequality

$$b(\alpha) = b + 2b_2\alpha - 3b_3\alpha^2 < 1 \tag{9}$$

distinguishes the condition for the stability of the equilibrium state $u(t, x) = \alpha$ of the main boundary value problem (1), (2).

For example, the equilibrium state $u(t, x) = 0$ of the main boundary value problem (1), (2) is stable, if $b < 1$. Moreover, for $b < 1$ and sufficiently small b_2^2 (for example, $b_2 = 0$) all equilibrium states $u(t, x) = \alpha$ of the boundary value problem (1), (2) are stable. It is worth to emphasize, that the equilibrium state $u(t, x) = \alpha$ cannot be asymptotically stable since the boundary value problem (1), (2) has a family of equilibrium states $u(t, x) = b$, where b is an arbitrary real constant.

Remark 2. If we consider the nonlinear boundary value problem (5), (6), then for $b(\alpha) < 1$ and $b_2(\alpha) = 0$ the following statement holds.

Lemma 3. *Let* $v(t, x)$ *be the solution which exists for any* $t \geq 0$. *Then*

$$\lim_{t\to\infty} \int_0^{2\pi} v^2(t, x)dx = 0.$$

Indeed, if we multiply Eq. (5) by $v(t,x)$ and integrate the resulting equality from 0 to 2π, then after transformations of the integrals on its right-hand side we obtain the following equality

$$\frac{1}{2}\frac{d}{dt}\int\limits_0^{2\pi} v^2 dx = -\int\limits_0^{2\pi} v_{xx}^2 dx + b(\alpha)\int\limits_0^{2\pi} v_x^2 dx - 3b_3\int\limits_0^{2\pi} v^2 v_x^2 dx.$$

The right hand side of the last equality is negative. Recall, that $\int\limits_0^{2\pi} v_x^2 dx \leq \int\limits_0^{2\pi} v_{xx}^2 dx$. The last inequality can be proved using the Parseval's identity.

A similar statement holds if

$$b_2^2(\alpha) - 3(1 - b(\alpha))b_3 < 0.$$

3 Local Bifurcations of Auxiliary Boundary Value Problems

In this section, we consider the question of local bifurcations for the auxiliary boundary value problem (5), (6) . Let $b(\alpha) = 1 + \gamma\varepsilon$, where $\varepsilon \in (0, \varepsilon_0)$, $0 < \varepsilon_0 << 1$, i.e. ε is a small positive parameter, $\gamma = \pm 1$ and the corresponding version will be selected later in the bifurcation analysis of the boundary value problem (5), (6).

Let $b_3 > 0$. First, the question arises about the possibility of implementing the equality $b(\alpha) = 1 + \gamma\varepsilon$. Of course, it is related to the possibility of implementing the equality $b(\alpha) = 1$, which comes down to the analysis of the quadratic equation

$$3b_3\alpha^2 - 2b_2\alpha + 1 - b = 0. \tag{10}$$

This equation has two real roots $\alpha_{1,2} = \dfrac{b_2 \pm \sqrt{D}}{3b_3}$, if $D = b_2^2 + 3b_3(b-1) > 0$. For $D < 0$ the main boundary value problem (1), (2) lacks equilibrium states $u(t,x) = \alpha$, for which there is a critical case in the stability problem for this equilibrium state. Moreover, for $D < 0$ all the homogeneous equilibrium states $u(t,x)$ of the boundary value problem (1), (2) are stable (for example, $D < 0$, if $b < 1, b_2 = 0$).

So, let $D > 0$. Then, there is an equality $b(\alpha) = 1 + \gamma\varepsilon$ for $\alpha_1(\varepsilon) = \alpha_1 + \beta_1(\varepsilon)$ or $\alpha_2(\varepsilon) = \alpha_2 + \beta_2(\varepsilon)$, where α_1, α_2 are roots of the quadratic equation (10) and $\beta_j(\varepsilon)(j = 1, 2)$ are analytic functions of ε. For these, the equalities $\beta_j(0) = 0$, $\beta_j(\varepsilon) = \beta_j\varepsilon + o(\varepsilon)$, where $\beta_j = -\gamma/(6b_3\alpha_j - 2b_2)$, are satisfied. It is clear that $6b_3\alpha_j - 2b_2 \neq 0$, since the roots α_j are simple.

A critical case arises if $b_3 = 0$ or $D = 0$. For $b_3 = 0$ the equality $b(\alpha) = 1 + \gamma\varepsilon$ holds, if

$$\alpha = \alpha_1 + \beta_1\varepsilon,\ \alpha_1 = \frac{1-b}{2b_2},\ \beta_1 = \frac{\gamma}{2b_2}.$$

In this work we shall not consider the special case $D = 0$.

As a result, the auxiliary boundary value problem (5), (6) can be rewritten as follows

$$v_t = A(\varepsilon)v + c(v^2)_x - b_2(\varepsilon)(v^2)_{xx} + b_3(v^3)_{xx}, \tag{11}$$

$$v(t, x + 2\pi) = v(t, x),\ M_0(v) = 0, \tag{12}$$

where $b_2(\varepsilon) = b_{4j} - 3b_3\beta_j\varepsilon + o(\varepsilon)$, and $b_{4j} = b_2 - 3b_3\alpha_j$.

Finally,

$$A(\varepsilon)v = -v_{xxxx} - (1 + \gamma\varepsilon)v_{xx} + c(\varepsilon)v_x,\ c(\varepsilon) = 2c\alpha_j + 2c\beta_j\varepsilon + o(\varepsilon),\ j = 1, 2.$$

Next, we will use a shortened version of the notation, considering the index j already fixed, and set $c(0) = \sigma$ $(c(0) = \sigma_j = 2c\alpha_j)$, $2c\beta_j = \delta$.

The LDO $A(\varepsilon)$ has two eigenvalues $\lambda_{1,2}(\varepsilon) = \gamma\varepsilon \pm i(\sigma + \delta\varepsilon + o(\varepsilon))$. The rest of its eigenvalues lie in the half-plane of the complex plane divided by the inequality $Re\lambda \leq -\gamma_0 < 0$. Finally, $Re\lambda_{1,2}(\varepsilon) = \gamma\varepsilon$, i.e. $Re\lambda'_{1,2}(\varepsilon)|_{\varepsilon=0} = \gamma \neq 0$, $Im\lambda_{1,2}(0) = \sigma \neq 0$, if $c \neq 0, \alpha_j \neq 0\,(b \neq 1)$.

We point out that the first group of conditions of the Andronov–Hopf theorem is satisfied (see, for example, [11]).

We turn to the second part of this theorem and state it in a modern form, which corresponds to the problem under study. It follows from the results of [11, 12] that the nonlinear boundary value problem (11), (12) has a smooth two-dimensional invariant manifold $M_2(\varepsilon, \alpha)$ (in different terminology $M_2(\varepsilon, \alpha)$ is a central manifold) in a neighborhood of the zero equilibrium state. All solutions belonging to this neighborhood tend to $M_2(\varepsilon, \alpha)$.

The analysis of the dynamics of solutions to the boundary value problem (11), (12) can be reduced to studying a system of two ordinary differential equations. In complex form, this system can be written as

$$\dot{z} = \varepsilon[(a_1 + ia_2)z + (l_1 + il_2)z|z|^2] + o(\varepsilon), \tag{13}$$

where $z = z(t) = z_1(t) + iz_2(t)$, $a_1, a_2, l_2, l_2 \in R$, the value l_1 is called the first Lyapunov value. If $l_1 \neq 0$, then, instead of Eq. (13), which is usually called the normal Poincaré form, consider its shortened version ("truncated normal form")

$$z' = \varepsilon[(a_1 + ia_2)z + (l_1 + il_2)z|z|^2]. \tag{14}$$

To construct the right-hand side of differential equation (14), it is possible and convenient to apply the following algorithm [13–17], which can be interpreted as an adaptation of the well-known Krylov-Bogoliubov method for partial differential equations

We now consider the nonlinear boundary value problem (11), (12). In its analysis, the notation $A(0)v = A_0 v$, $B_0 v = -\gamma v_{xx} + \delta v_x$ will be used. Recall that $b_2(\varepsilon) = b_{4j} + O(\varepsilon)$. Next we will use a simplified version of the notation $b_2(\varepsilon) = b_4 + O(\varepsilon)$, assuming, that α_j (root of equation (10)) is chosen. Finally, the LDO $A(0)$ has a pair of purely imaginary eigenvalues $\pm i\sigma$.

The solutions of the boundary value problem (11), (12) belonging to $M_2(\alpha, \varepsilon)$ should be sought in the next form

$$v(t, x, \varepsilon) = \varepsilon^{1/2} v_1(t, x, z, \overline{z}) + \varepsilon v_2(t, x, z, \overline{z}) + \varepsilon^{3/2} v_3(t, x, z, \overline{z}) + o(\varepsilon^{3/2}), \quad (15)$$

where the functions v_1, v_2, v_3 possess the following properties:

1) $v_1(t, x, z, \overline{z}) = z \exp(ix + i\sigma t) + \overline{z} \exp(-ix - i\sigma t)$, where $z = z(t)$ are normal form solutions, i.e. the derivative of the complex-valued function is calculated by virtue of Eq. (14).
2) The functions v_2, v_3 depend quite smoothly on their variables. In particular, for fixed, for fixed $t, z, \overline{z}\, v_j(t, x, z, \overline{z}) \in H^4_{2,0}$, if we consider them as functions of x We assume that they have the period $2\pi/\sigma$ with respect to t and, for them, the following equalities

$$M_{\pm}(v_j) = 0,\ j = 2, 3,\ M_{\pm}(v_j) = \frac{1}{2\pi} \int_0^{2\pi} v_j \exp(\pm i\sigma t) \exp(\pm ix) dx dt = 0.$$

hold. This class of solutions will be denoted by V.

We substitute the sum (15) into the boundary value problem (11), (12) and equate the obtained expressions for $\varepsilon, \varepsilon^{3/2}$. As a result, we obtain linear nonhomogeneous boundary value problems for determining v_2, v_3 :

$$v_{2t} - A_0 v_2 = F_2(t, x), \quad (16)$$

$$v_2(t, x + 2\pi) = v_2(t, x),\ M_0(v_2) = 0, \quad (17)$$

$$v_{3t} - A_0 v_3 = F_3(t, x), \quad (18)$$

$$v_3(t, x + 2\pi) = v_3(t, x),\ M_0(v_3) = 0. \quad (19)$$

Here,

$$F_2(t, x) = c(v_1^2)_x - b_4(v_1^2)_{xx},$$
$$F_3(t, x) = -[\psi \exp(i\sigma t + ix) + \overline{\psi} \exp(-i\sigma t - ix)] + B_0 v_1$$
$$+2c(v_1 v_2)_x - 2b_4(v_1 v_2)_{xx} + b_3(v_1^3)_{xx},$$

and ψ denotes the normalized right-hand side of the shortened normal form (14). We emphasize once again that the derivative with respect to the variable t is calculates in view of equation (14).

The boundary value problem (16), (17) is uniquely correctly solvable in the class of functions V. Indeed, the conditions for its solvability in this class of functions

$$M_{\pm}(F_2) = 0$$

are satisfied. The equations $M_{\pm}(v_2) = 0$ distinguish the needed solution. In our case,

$$v_2(t, x, z) = \eta_2 z^2 \exp(2i\sigma t + 2ix) + \overline{\eta}_2 \overline{z^2} \exp(-2i\sigma t - 2ix),$$

where the complex constant η_2 has the following form

$$\eta_2 = \frac{2b_4 + ci}{6}.$$

We now turn to the analysis of the linear inhomogeneous boundary value problem (18), (19). The use of solvability conditions allows us to define ψ:

$$\psi = (a_1 + ia_2)z + (l_1 + il_2)z|z|^2,$$

where

$$a_1 = \gamma,\ a_2 = \delta,\ l_1 = -(3b_3 + \frac{c^2}{3}) + \frac{2}{3}b_4^2,$$
$$l_2 = cb_4,\ b_4 = b_2 - 3b_3\alpha,\ \alpha = \alpha_j,\ j = 1, 2.$$

We point out once again that $\delta = \delta_j = 2c\beta_j$, $b_4 = b_{4j} = b_2 - 3b_3\alpha_j$, $j = 1, 2$, i.e. These quantities depend on the choice of the roots of the quadratic equation (10).

Let $c \neq 0$. The case $c = 0$ is considered separately. Let also $l_1 \neq 0$. Then, the following holds.

Lemma 2. *The differential equation (14) has a family of periodic solutions*

$$z(t) = \rho_0 \exp(i\varepsilon\omega t + i\varphi_0), \tag{20}$$

where φ_0 is an arbitrary real value.

Here,

$$\rho_0 = \sqrt{-\gamma/l_1}, \omega = \delta - l_2\gamma/l_1.$$

For this $\gamma = 1$, if $l_1 < 0$ and $\gamma = -1$, if $l_1 > 0$.

The family of periodic solutions (20) generates the limit cycle $C(\varepsilon)$ of differential equation (14). This limit cycle is stable (local attractor), if $l_1 < 0$ and this cycle is unstable if $l_1 > 0$.

To prove the last statement, we can set

$$z(t) = \rho(t)\exp(i\varphi(t)).$$

Then, instead of Eq. (14), we obtain the system

$$\dot{\rho} = \varepsilon(\gamma\rho + l_1\rho^3),\ \rho = \rho(t) \geq 0,\ \dot{\varphi} = \varepsilon(\delta + l_2\rho^2).$$

The first equation of this system has a zero equilibrium state which is unstable for $\gamma = 1$. For such γ it has an asymptotically stable equilibrium $\rho(t) = \rho_0 = \sqrt{-\gamma/l_1}$, if $l_1 < 0$.

If $\gamma = -1$, then the first equation of the last system has an unstable equilibrium state $\rho(t) = \rho_0 = \sqrt{-\gamma/l_1}$, if, of course, $l_1 > 0$. Obviously, for $\gamma = -1$ the zero solution of this equation is asymptotically stable

The validity of the statement follows from the results of [18, 19].

Theorem 1. *There exists $\varepsilon_0 > 0$, such that for all $\varepsilon \in (0, \varepsilon_0)$ the boundary value problem (11), (12) has the limit cycle $C(\alpha, \varepsilon)$ corresponding to the limit cycle $C(\varepsilon)$ of the normal form (14). The cycle $C(\alpha, \varepsilon)$ inherits the stability of the cycle $C(\varepsilon)$. For the solutions which form this limit cycle, we have the following asymptotic formula*

$$\begin{gathered} v(t, x, \alpha, \varepsilon) = \varepsilon^{1/2}\rho_0[\exp(i\sigma(\varepsilon)t + ix + i\varphi_0) \\ + \exp(-i\sigma(\varepsilon)t - ix - i\varphi_0)] + \varepsilon\rho_0^2[\eta_2\exp(2i\sigma(\varepsilon)t + 2ix + 2i\varphi_0) \\ + \overline{\eta}_2\exp(-2i\sigma(\varepsilon)t - 2ix - 2i\varphi_0)] + O(\varepsilon^{3/2}), \end{gathered} \tag{21}$$

where φ_0 is an arbitrary constant, $\sigma(\varepsilon) = \sigma + \varepsilon\omega$, and the constants $\sigma, \omega \neq 0$ were indicated earlier.

We obtain the asymptotic formula (21) after substitution of solution (20) into the formula for solutions defining $M_2(\varepsilon)$ in parametric form, i.e. into equality (15). The asymptotic formula (21) can be rewritten in real form

$$\begin{gathered} v(t, x, \alpha, \varepsilon) = 2\varepsilon^{1/2}\rho_0\cos(\sigma(\varepsilon)t + x + \varphi_0) \\ + 2\varepsilon\rho_0^2[\eta_{21}\cos(2\sigma(\varepsilon)t + 2x + 2\varphi_0) \\ - \eta_{22}\sin(2\sigma(\varepsilon)t + 2x + 2\varphi_0)] + O(\varepsilon^{3/2}), \end{gathered} \tag{22}$$

where $\eta_{21} = \dfrac{b_4}{3}$, $\eta_{22} = \dfrac{c}{6}$.

Remark 3. For $c = 0$ we obtain $\sigma = 0, l_2 = 0$. In this case, the normal form (14) has equilibrium states

$$z(t) = \rho\exp(i\varphi),\ \varphi \in R,\ \ \rho \in R_+.$$

Hence the nonlinear boundary value problem (11), (12) has the family of equilibrium states corresponding to the equilibrium states of the normal form (14).

4 The Main Result

We now return to the analysis of the main boundary value problem. Periodic solutions of the main boundary value problems (1), (2), correspond to the periodic solutions (23) of the auxiliary boundary value problem (5), (6), if, of course, $b(\alpha) = 1 + \gamma\varepsilon, \sigma \neq 0, \alpha \neq 0$.

$$\begin{aligned} u(t, x, \alpha(\varepsilon), \varepsilon) &= \alpha(\varepsilon) + v(t, x, \alpha(\varepsilon), \varepsilon) \\ &= \alpha(\varepsilon) + 2\varepsilon^{1/2}\rho_0 \cos(\sigma(\varepsilon)t + x + \varphi_0) \\ &+ 2\varepsilon\rho_0^2(\eta_{21} \cos(2\sigma(\varepsilon)t + 2x + 2\varphi_0) \\ &- \eta_{22} \sin(2\sigma(\varepsilon)t + 2x + 2\varphi_0)) + O(\varepsilon^{3/2}). \end{aligned} \tag{23}$$

The family of solutions (23) in the phase space of solutions of the boundary value problem (1), (2) generates a limit cycle for each $\varepsilon \in (0, \varepsilon_0)$ and all those α, for which $b(\alpha) = 1 + \gamma\varepsilon$, where $\alpha = \alpha(\varepsilon)$ is one of the roots of the equation

$$3b_3\alpha^2 - 2b_2\alpha + (1 + \gamma\varepsilon - b) = 0 \tag{24}$$

for $\varepsilon \in (0, \varepsilon_0)$, i.e. $\alpha(\varepsilon) \in I(\varepsilon)$, where $I = (\alpha_j, \alpha_j(\varepsilon_0))$ or $I = (\alpha_j(\varepsilon_0), \alpha_j)$, and $\alpha_j(\varepsilon_0)$ is a corresponding root of the quadratic equation (24) for $\varepsilon = \varepsilon_0$, and α_j one of the roots of Eq. (24). There are two such intervals in the general case $(j = 1, 2)$.

Let the equilibrium state of the boundary value problem (1), (2) $u(t, x) = \alpha$ or $\alpha \in I_j(\varepsilon)$ be chosen. Then, as already noted, this boundary value problem has the limit cycle $C(\alpha, \varepsilon)$ and for $l_1 < 0$ it is attractive in the following sense. Let $u(t, x, \alpha)$ be some solution from its neighborhood and, in addition, $u(t, x, \alpha) \in H_4(\alpha)$. Then, over time it approaches to $C(\alpha, \varepsilon)$. This fact follows from the analysis of the auxiliary boundary value problem (11), (12). If the limit cycles $C(\varepsilon)$ of the auxiliary boundary value problem (11), (12) are unstable for the considered ε, then the solutions $u(t, x) \in H_4(\alpha)$ leave the neighborhood of the corresponding limit cycle $C(\alpha, \varepsilon)$.

Let $V_j = \underset{\alpha \in I_j(\varepsilon)}{\cup} C(\alpha, \varepsilon)$, $j = 1, 2$. Then, the following statement holds.

Theorem 2. For all $\varepsilon \in (0, \varepsilon_0)$ an invariant two-parameter manifold for the solutions of the boundary value problem (1), (2) has the following properties:

- for $l_1 < 0$ the manifold V_j is a local attractor;
- for $l_1 > 0$ the manifold V_j is a saddle invariant set;
- the variety V_j is formed by t by periodic solutions (23), the period of which depends on the choice of $\varepsilon \in (0, \varepsilon_0)$;
- solutions of the two-parameter family (23) are unstable in the Lyapunov sense in the metric of the phase space of solutions of the boundary value problem (1), (2).

It follows from the previous constructions that it remains to verify only the last part of Theorem 2 for $l_1 < 0$, i.e. in the case when the limit cycles $C(\varepsilon)$ of the auxiliary boundary value problem are stable.

Let $\varepsilon = \varepsilon_* \in (0, \varepsilon_0)$ and $u_*(t, x)$ be a solution of the family (23) for chosen φ_*. We now set $\varepsilon_\mu = \varepsilon_*(1 + \mu)$ and consider the solutions of family (23) for such a chosen ε and $\varphi = \varphi_*$. We denote such solutions by $u_\mu(t, x)$. It is obvious that

$$\lim_{\mu\to 0} ||u_\mu(0, x) - u_*(0, x)||_{H_4} = 0$$

due to the continuous dependence of the solutions on the parameter ε. But, there is a sequence $t_k = t_k(\varepsilon_*, \mu)$, such that $||u_\mu(t_k, x) - u_*(t_k, x)||_{H_4} \geq r$, where $r = r(\varepsilon_*) > 0$, but do not depend on μ. Moreover, when checking the last inequality, it suffices to restrict ourselves to considering the "main" parts of the corresponding asymptotic formulas and show that the inequality

$$||w_\mu(t_k, x) - w_*(t_k, x)||_{H_4} \geq 2r,$$

where

$$w_\mu = \alpha(\varepsilon_\mu) + 2\varepsilon_\mu^{1/2}\rho_0 \cos(\sigma(\varepsilon_\mu)t_k + x + \varphi_*),$$
$$w_* = \alpha(\varepsilon_*) + 2\varepsilon_*^{1/2}\rho_0 \cos(\sigma(\varepsilon_*)t_k + x + \varphi_*),$$

where $\rho_0 = \sqrt{-\gamma/l_1}$ ($\gamma = 1, l_1 < 0$ and it does not depend on ε and μ), $\sigma(\varepsilon) = \sigma + \varepsilon\omega$ and the constant (which was indicated earlier) $\omega \neq 0$, $\varepsilon_\mu = \varepsilon_*(1 + \mu)$.

Obviously, for any function $g(x) \in H_4$ the following inequality holds

$$||g||_{H_4} \geq ||g_{xxxx}||_{L_2(0,2\pi)}.$$

Therefore, it is enough to verify the inequality

$$||w_{\mu_{xxxx}} - w_{*_{xxxx}}||_{L_2(0,2\pi)} \geq 2r$$

for some $t = t_k$.

In turn, it will be satisfied if we specify t_k such that

$$||y||_{L_2(0,2\pi)} \geq 4r,$$

where $y = y(t_k, x) = 2\varepsilon_*^{1/2}\rho_0[\cos(\sigma(\varepsilon_\mu)t_k + x + \varphi_*) - \cos(\sigma(\varepsilon_*)t_k + x + \varphi_*)]$.

It can readily be shown that

$$\int_0^{2\pi} y^2(t_k, x)dx = 16\varepsilon_*\rho_0^2\pi \sin^2(\frac{\omega\mu\varepsilon_* t_k}{2})$$

and, consequently, for

$$t_k = \frac{1}{\mu\omega\varepsilon_*}(\pi + 2\pi k),$$

where $k = 1, 2, 3, \ldots,$ if $\mu\omega > 0$ and $k = -1, -2, -3, \ldots,$ if $\mu\omega < 0$. The given integral is equal to $16\varepsilon_* \rho_0^2 \pi$, i.e. as $r = r(\varepsilon_*)$ we can choose $\varepsilon_*^{1/2}\rho_0 =$ $= \varepsilon_*^{1/2}(-\gamma/l_1)^{1/2}(\pi)^{1/2} > 0$. So, instability of solutions of family (23) is proved.

5 Conclusion

In this work, the existence of a local attractor with unstable periodic solutions is shown. Therefore, 2 out of 3 points of the definition of a chaotic attractor according to the definition of Devaney [20] are fulfilled. The attractor found does not satisfy only the third point of this definition, since the ergodicity of the flow generated by the considered boundary value problem is absent on the local attractor.

The reported study was funded by RFBR according to the research project No 18-01-00672.

References

1. Cahn, J.W., Hilliard, J.E.: Free energy of a nonuniform system. I. Interfacial free energy. J. Chem. Phys. **28**, 258–267 (1958)
2. Podolny, A., Zaks, M.A., Rubinstein, B.Y., Golovin, A.A., Nepomnyashchy, A.A.: Dynamics of domain walls governed by the convective Cahn-Hilliard equation. Phys. D **201**, 91–305 (2005)
3. Frolovskaya, O.A., Pukhnachev, V.V.: Stationary solutions of quadratic Cahn–Hilliard equation and their stability. In: AIP Conference Proceedings, vol. 1561, pp. 47–52 (2013)
4. Frolovskaya, O.A., Admaev, O.V., Pukhnachev, V.V.: Special case of the Cahn-Hilliard equation. Siber. Electon. Math. Rep. **10**, 324–334 (2013)
5. Novick-Cohen, A., Segel, L.A.: Nonlinear aspects of the Cahn-Hilliard equation. Phys. D **10**, 277–298 (1984)
6. Kulikov, A.N., Kulikov, D.A.: Local bifurcations in the Cahn-Hilliard and Kuramoto-Sivashinsky equations and in their generalizations. Comput. Math. Math. Phys. **59**, 630–643 (2019)
7. Kulikov, A.N., Kulikov, D.A.: Spatially ingomogeneous solutions in two boundary value problems for the Cahn-Hilliard equations. Belgorod State Univ. Sci. Bull. Math. Phys. **51**, 21–32 (2019)
8. Sobolevskii, P.E.: Equations of a parabolic type in a Banach space. Moscov. Mat. Obsc. **10**, 297–350 (1961)
9. Lions, J.L., Magenes, E.: Problemes aux limit es nonhomogenes et applications, vol. 1. Dunod, Paris (1968)
10. Krein, S.G.: Linear Equations in Banach Spaces. Springer, New York (1982)
11. Marsden, J.E., McCraken, M.: The Hopf Bifurcations and its Applications. Springer, New York (1976)
12. Kulikov, A.N.: Inertial manifolds of nonlinear self-oscillations of differential equations in a Hilbert space. Preprint 85 of Institute of M.V. Keldysh applied mathematics, Moscow (1991)
13. Kulikov, A.N., Kulikov, D.A.: Formation of wavy nanostructures on the surface of flat substrates by ion bombardment. Comput. Math. Math. Phys. **52**, 930–945 (2012)
14. Kulikov, A., Kulikov, D.: Bifurcation in Kuramoto-Sivashinsky equation. Pliska Stud. Math. **25**, 101–110 (2015)

15. Kulikov, A.N., Kulikov, D.A.: Local bifurcations in the periodic boundary value problem for the generalized Kuramoto-Sivashinsky. Autom. Remote Control **78**, 1955–1966 (2017)
16. Kulikov, A.N., Kulikov, D.A.: Bifurcations in a boundary value problem of nanoelectronics. J. Math. Sci. **208**, 211–221 (2015)
17. Kulikov, A.N., Kulikov, D.A.: Spatially inhomogeneous solutions for a modified Kuramoto-Sivashinsky equation. J. Math. Sci. **219**, 173–183 (2016)
18. Kolesov, A.Y., Kulikov, A.N., Rozov, N.H.: Invariant tori of a class of point mapping: the annulus principle. Differ. Equ. **39**, 614–631 (2003)
19. Kolesov, A.Y., Kulikov, A.N., Rozov, N.K.: Invariant tori of a class of point transformations: preservation of an invariant torus under perturbations. Differ. Equ. **39**, 775–790 (2003)
20. Devaney, R.L.: An Introduction to Chaotic Dynamical Systems. Westview Press, Colorado (1989)

The Numerical Solution of Wave Equation with Delay for the Case of Variable Velocity Coefficient

Ekaterina Tashirova

Abstract The wave equations with delay and variable velocity coefficient are considered. A family of grid methods is constructed for the numerical solution of this equations. The convergence of the constructed method is investigated by means of embedding into a general difference scheme with delay. Results of calculating test examples are presented.

1 Problem Statement

Let us consider wave equation with delay

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(k(x,t)\frac{\partial u}{\partial x}\right) + f(x,t,u(x,t),u_t(x,\cdot))\colon t_0 \le t \le T,\ 0 \le x \le X, \tag{1}$$

with boundary conditions

$$u(0,t) = g_1(t),\ u(X,t) = g_2(t)\colon t_0 \le t \le T \tag{2}$$

and initial conditions

$$u(x,t) = \varphi(x,t)\colon 0 \le x \le X,\ t_0 - \tau \le t < t_0, \tag{3}$$

Here, $u(x,t)$ is the required function; $u_t(x,\cdot) = \{u(x,t+\xi), -\tau \le \xi < 0\}$—is the history function of the required function to the moment t; τ is the value of delay; $f(x,t,u(x,t),u_t(x,\cdot))$—is a functional defined on $[0,X] \times [t_0,T] \times R \times Q[-\tau,0)$, $Q = Q[-\tau,0)$—is the set of piece-wise continuous functions on $[-\tau,0)$ with a finite number of discontinuity points of the first kind and continuous on the right at the discontinuity points; $\|u(\cdot)\|_Q = \sup_{\xi\in[-\tau,0)}|u(\xi)|$.

E. Tashirova (✉)
Ural Federal University, Yekaterinburg, Russia
e-mail: linetisa@yandex.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_15

We assume that function $k(x,t)$ is continuous and satisfies the following conditions

$$0 < c_1 \leq k(x,t) \leq c_2 \colon t_0 \leq t \leq T,\ 0 \leq x \leq X; \tag{4}$$

$$\frac{|k(x,t) - k(x,t-\Delta)|}{\Delta} \leq c_3 k(x,t-\Delta), t > t_0. \tag{5}$$

We also assume that the functional $f(x,t,u,u(\cdot))$ is Lipschitz with respect to the last two arguments, i.e. there exists a constant L_f such that, for all $x \in [0,X]$, $t \in [t_0,T]$, $u^1 \in R^1$, $u^2 \in R^1$, $v^1(\cdot) \in Q[-\tau,0)$, $v^2(\cdot) \in Q[-\tau,0)$ the following inequality holds:

$$|f(x,t,u^1,v^1(\cdot)) - f(x,t,u^2,v^2(\cdot))|$$

$$\leq L_f(|u^1 - u^2| + \|v^1(\cdot) - v^2(\cdot)\|_{Q[-\tau,0)}). \tag{6}$$

In addition we assume that the functional f and functions g_1, g_2, φ, k are such that the problem has a unique solution $u(x,t)$ [4].

2 Difference Method

Let us divide the interval $[0,X]$ into parts with step $h = X/N$.

$$x_i = ih,\ i = 0,1\ldots,N.$$

Let us divide the interval $[t_0,T]$ into parts with step Δ. $m = \tau/\Delta$ is an integer.

$$t_j = t_0 + j\Delta,\ j = -m,\ldots,M.$$

We will denote approximations of the exact solution of $u(x_i,t_j)$ by u^i_j.

Introduce the discrete history to the moment t_j for each fixed i:

$$\{u^i_k\}_j = \{u^i_k \colon j - m \leq k \leq j\}.$$

Mapping I

$$I : \{u^i_k\}_j \to v^i_j(\cdot) \in Q[-\tau,\Delta]$$

will be called an interpolation-extrapolation operator for the discrete history.

We will say that the interpolation–extrapolation operator has order of error p on the exact solution if constants C_1 and C_2 exist and they are such that for all $i = 0,\ldots,N$, $j = 0,\ldots,M$ and $t \in [t_j - \tau, t_{j+1}]$ the following inequality holds

$$|v^i_j(t) - u(x_i,t)| \leq C_1 \max_{j-m \leq k \leq j} |u^i_k - u(x_i,t_k)| + C_2\Delta^p.$$

The piece-wise linear interpolation

$$v_j^i(\xi) = \frac{1}{\Delta}((t_l - t_j - \xi)u_{l-1}^i + (t_j + \xi - t_{l-1})u_l^i),\ t_{l-1} \le t_j + \xi \le t_l,\ -\tau \le \xi \le 0 \tag{7}$$

has second order.

Consider a family of methods with weight ($0 \le s \le 1$):

$$\begin{aligned}\frac{u_{j+1}^i - 2u_j^i + u_{j-1}^i}{\Delta^2} &= s\left(k_j^{i+1/2}\frac{u_{j+1}^{i+1} - u_{j+1}^i}{h^2} - k_j^{i-1/2}\frac{u_{j+1}^i - u_{j+1}^{i-1}}{h^2}\right)\\ &+ s\left(k_j^{i+1/2}\frac{u_{j-1}^{i+1} - u_{j-1}^i}{h^2} - k_j^{i-1/2}\frac{u_{j-1}^i - u_{j-1}^{i-1}}{h^2}\right)\\ &+ (1-2s)\left(k_j^{i+1/2}\frac{u_j^{i+1} - u_j^i}{h^2} - k_j^{i-1/2}\frac{u_j^i - u_j^{i-1}}{h^2}\right) + F_j^i(v_j^i(\cdot)),\\ &i = 1, \dots N-1,\ j = 0, \dots M-1\end{aligned} \tag{8}$$

with boundary conditions

$$u_j^0 = g_1(t_j),\ u_j^N = g_2(t_j)$$

and initial conditions

$$u_j^i = \varphi(x_i, t_j):\ -m \le j \le 0,$$

where $k_j^{i+1/2} = k(x_i + h/2, t_j)$, $k_j^{i-1/2} = k(x_i - h/2, t_j)$; $F_j^i(v(\cdot))$—is a functional defined on $v(\cdot) = v_j^i(\cdot) = I(\{u_k^i\}_j) \in Q[-\tau, \Delta]$, and connected with the functional $f(x_i, t_j, u_j^i, v_j^i(\cdot))$; also we assume that functional $F_j^i(v(\cdot))$ is Lipschitz with respect to the variable $v(\cdot)$ with the constant L_F.

For $s = 0$, we obtain an explicit scheme. For other s, $0 < s \le 1$, for each fixed j, the system is linear tridiagonal with respect to u_{j+1}^i with diagonal dominance; hence, it can be effectively solved by the tridiagonal matrix algorithm.

The residual (without interpolation) of the method is the value:

$$\begin{aligned}\psi_j^i = &\frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{\Delta^2} -\\ &- s\Big(k(x_i + h/2, t_j)\frac{u(x_{i+1}, t_{j+1}) - u(x_i, t_{j+1})}{h^2} -\\ &- k(x_i - h/2, t_j)\frac{u(x_i, t_{j+1}) - u(x_{i-1}, t_{j+1})}{h^2}\Big) +\\ &- s\Big(k(x_i + h/2, t_j)\frac{u(x_{i+1}, t_{j-1}) - u(x_i, t_{j-1})}{h^2} -\end{aligned}$$

$$- k(x_i - h/2, t_j)\frac{u(x_i, t_{j-1}) - u(x_{i-1}, t_{j-1})}{h^2}\Big) +$$
$$- (1-2s)\Big(k(x_i + h/2, t_j)\frac{u(x_{i+1}, t_j) - u(x_i, t_j)}{h^2} -$$
$$- k(x_i - h/2, t_j)\frac{u(x_i, t_j) - u(x_{i-1}, t_j)}{h^2}\Big) - F_j^i(u_{t_j}(x_i, \cdot)). \tag{9}$$

We will say that the residual has order $h^{p_1} + \Delta^{p_2}$, if there exists constant C such that $|\psi_j^i| \le C(h^{p_1} + \Delta^{p_2})$ for all $i = 1, \dots N-1,\ j = 0, \dots M-1$.

Theorem 1. *Suppose that the exact solution of the problem (1)–(3) has continuous partial derivatives up to the fourth order, function $k(x,t)$ has continuous partial derivatives with respect to t up to the third order and $F_j^i(v_j^i(\cdot)) = f(t_j, x_i, u_j^i, v_j^i(\cdot))$. Then for every $0 \le s \le 1$ the residual has order $h^2 + \Delta^2$.*

Proof. The residual of the method is the value:

$$\psi_j^i = \frac{u(x_i, t_{j+1}) - 2u(x_i, t_j) + u(x_i, t_{j-1})}{\Delta^2} -$$
$$- s\Big(k(x_i + h/2, t_j)\frac{u(x_{i+1}, t_{j+1}) - u(x_i, t_{j+1})}{h^2} -$$
$$- k(x_i - h/2, t_j)\frac{u(x_i, t_{j+1}) - u(x_{i-1}, t_{j+1})}{h^2}\Big) +$$
$$- s\Big(k(x_i + h/2, t_j)\frac{u(x_{i+1}, t_{j-1}) - u(x_i, t_{j-1})}{h^2} -$$
$$- k(x_i - h/2, t_j)\frac{u(x_i, t_{j-1}) - u(x_{i-1}, t_{j-1})}{h^2}\Big) +$$
$$- (1-2s)\Big(k(x_i + h/2, t_j)\frac{u(x_{i+1}, t_j) - u(x_i, t_j)}{h^2} -$$
$$- k(x_i - h/2, t_j)\frac{u(x_i, t_j) - u(x_{i-1}, t_j)}{h^2}\Big) - F_j^i(u_{t_j}(x_i, \cdot)).$$

Let us write the Taylor expansion of functions $u(x,t)$ and $k(x,t)$ in the neighborhood of points (x_i, t_j), (x_i, t_{j+1}), (x_i, t_{j-1})

$$u(x_i, t_{j+1}) = u(x_i, t_j) + \frac{\partial u}{\partial t}(x_i, t_j)\Delta + \frac{1}{2}\frac{\partial^2 u}{\partial t^2}(x_i, t_j)\Delta^2 + \frac{1}{6}\frac{\partial^3 u}{\partial t^3}(x_i, t_j)\Delta^3 + O(\Delta^4),$$

where $g = O(\Delta^4)$, if there exists constant C, such that the inequality $|g| \le C\Delta^4$ holds.

$$u(x_i, t_{j-1}) = u(x_i, t_j) - \frac{\partial u}{\partial t}(x_i, t_j)\Delta + \frac{1}{2}\frac{\partial^2 u}{\partial t^2}(x_i, t_j)\Delta^2 - \frac{1}{6}\frac{\partial^3 u}{\partial t^3}(x_i, t_j)\Delta^3 + O(\Delta^4),$$

$$u(x_{i-1},t_j)=u(x_i,t_j)-\frac{\partial u}{\partial x}(x_i,t_j)h+\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_j)h^2-\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_j)h^3+O(h^4),$$

$$u(x_{i+1},t_j)=u(x_i,t_j)+\frac{\partial u}{\partial x}(x_i,t_j)h+\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_j)h^2+\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_j)h^3+O(h^4),$$

$$\begin{aligned}u(x_{i-1},t_{j+1})&=u(x_i,t_{j+1})-\frac{\partial u}{\partial x}(x_i,t_{j+1})h+\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_{j+1})h^2\\&\quad-\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_{j+1})h^3+O(h^4),\end{aligned}$$

$$\begin{aligned}u(x_{i+1},t_{j+1})&=u(x_i,t_{j+1})+\frac{\partial u}{\partial x}(x_i,t_{j+1})h+\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_{j+1})h^2\\&\quad+\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_{j+1})h^3+O(h^4),\end{aligned}$$

$$\begin{aligned}u(x_{i-1},t_{j-1})&=u(x_i,t_{j-1})-\frac{\partial u}{\partial x}(x_i,t_{j-1})h+\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_{j-1})h^2\\&\quad-\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_{j-1})h^3+O(h^4),\end{aligned}$$

$$\begin{aligned}u(x_{i+1},t_{j-1})&=u(x_i,t_{j-1})+\frac{\partial u}{\partial x}(x_i,t_{j-1})h+\frac{1}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t_{j-1})h^2\\&\quad+\frac{1}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t_{j-1})h^3+O(h^4).\end{aligned}$$

$$k(x_i+h/2,t_j)=k(x_i,t_j)+\frac{h}{2}\frac{\partial k}{\partial x}(x_i,t_j)+O(h^2)$$

$$k(x_i-h/2,t_j)=k(x_i,t_j)-\frac{h}{2}\frac{\partial k}{\partial x}(x_i,t_j)+O(h^2)$$

Using the expansions above we obtain

$$\begin{aligned}\psi_j^i&=\frac{\partial^2 u}{\partial t^2}(x_i,t_j)+O(\Delta^2)\\&\quad-s\left(k(x_i,t_j)\frac{\partial^2 u}{\partial x^2}(x_i,t_{j+1})+\frac{\partial k}{\partial x}(x_i,t_j)\frac{\partial u}{\partial x}(x_i,t_{j+1})+O(h^2)\right)\\&\quad-s\left(k(x_i,t_j)\frac{\partial^2 u}{\partial x^2}(x_i,t_{j-1})+\frac{\partial k}{\partial x}(x_i,t_j)\frac{\partial u}{\partial x}(x_i,t_{j-1})+O(h^2)\right)\end{aligned}$$

$$- (1-2s)\left(k(x_i,t_j)\frac{\partial^2 u}{\partial x^2}(x_i,t_j)+\frac{\partial k}{\partial x}(x_i,t_j)\frac{\partial u}{\partial x}(x_i,t_j)+O(h^2)\right)$$
$$- f(t_j,x_i,u(x_i,t_j),u_{t_j}(x_i,\cdot)).$$

Let us write the Taylor expansion of functions $\frac{\partial^2 u}{\partial x^2}(x,t)$ and $\frac{\partial u}{\partial x}(x,t)$ in the neighborhood of points (x_i,t_j)

$$\frac{\partial^2 u}{\partial x^2}(x_i,t_{j+1})=\frac{\partial^2 u}{\partial x^2}(x_i,t_j)+\frac{\partial^3 u}{\partial t\partial x^2}(x_i,t_j)\Delta+O(\Delta^2),$$

$$\frac{\partial^2 u}{\partial x^2}(x_i,t_{j-1})=\frac{\partial^2 u}{\partial x^2}(x_i,t_j)-\frac{\partial^3 u}{\partial t\partial x^2}(x_i,t_j)\Delta+O(\Delta^2).$$

$$\frac{\partial u}{\partial x}(x_i,t_{j+1})=\frac{\partial u}{\partial x}(x_i,t_j)+\frac{\partial^2 u}{\partial t\partial x}(x_i,t_j)\Delta+O(\Delta^2),$$

$$\frac{\partial u}{\partial x}(x_i,t_{j-1})=\frac{\partial u}{\partial x}(x_i,t_j)-\frac{\partial^2 u}{\partial t\partial x}(x_i,t_j)\Delta+O(\Delta^2),$$

Then we obtain

$$\psi_j^i=\frac{\partial^2 u}{\partial t^2}(x_i,t_j)-k(x_i,t_j)\left(\frac{\partial^2 u}{\partial x^2}(x_i,t_j)\right)-\frac{\partial k}{\partial x}(x_i,t_j)\left(\frac{\partial u}{\partial x}(x_i,t_j)\right)$$
$$- f(t_j,x_i,u(x_i,t_j),u_{t_j}(x_i,\cdot))+O(\Delta^2+h^2).$$

According to (1) $\psi_j^i=O(\Delta^2+h^2)$. □

3 Convergence

Denote the error of the method at the nodes by $\varepsilon_j^i=u(x_i,t_j)-u_j^i$.

We say that the method converges with order $h^p+\Delta^q$, if there exists constant C independent of $\Delta,\ h$, such that the inequality $|\varepsilon_j^i|\le C(h^p+\Delta^q)$ holds for all $i=0,\dots,N,\ j=0,\dots M$.

Let us investigate the convergence of method (8) by means of embedding it into the general difference scheme with delay [1, 2].

For each t_j, denote the values of the discrete model by $\widetilde{\gamma}_j=(u_j^0,u_j^1,\dots u_j^N)^T\in\widetilde{\Gamma}$, where T is the transposition symbol, $\widetilde{\Gamma}$—is a vector space of dimension $N+1$ with the scalar product:

$$(\widetilde{\gamma},\widetilde{\omega})=\sum_{i=1}^{N-1}\widetilde{\gamma}^i\widetilde{\omega}^i h,\, \widetilde{\gamma}=(\widetilde{\gamma}^0,\widetilde{\gamma}^1,\ldots\widetilde{\gamma}^N)\in\widetilde{\Gamma},\ \widetilde{\omega}=(\widetilde{\omega}^0,\widetilde{\omega}^1,\ldots\widetilde{\omega}^N)\in\widetilde{\Gamma}. \quad (10)$$

and the norm

$$\|\widetilde{\gamma}_n\|_{\widetilde{\Gamma}}=\sqrt{(\widetilde{\gamma}_n,\widetilde{\gamma}_n)}. \quad (11)$$

In the space $\widetilde{\Gamma}$ we introduce operators $A(t)$, $R(t)$, $\widetilde{A}(t)$ and $\widetilde{R}(t)$

$$A(t)\widetilde{\gamma}_j=\widetilde{\mu}_j,\ \widetilde{\mu}_j=(\widetilde{\mu}_j^0,\widetilde{\mu}_j^1,\ldots,\widetilde{\mu}_j^N)^T,$$
$$\widetilde{\mu}_j^0=-k(x_i+h/2,t)\frac{u_j^i-u_j^{i-1}}{h^2}+k(x_i-h/2,t)\frac{u_j^{i+1}-u_j^i}{h^2},\ 1\le i\le N-1,$$
$$A(t)u_j^i=0,\ A(t)u_j^N=0, \quad (12)$$

$$\widetilde{A}(t)=\Delta^2A(t), \quad (13)$$

$$\widetilde{R}(t)=\Delta^2R(t)=\frac{1}{\Delta^2}E+s\widetilde{A}(t), \quad (14)$$

where E is the identity operator.

Then, system (8) can be written in the form

$$R(t_j)(\widetilde{\gamma}_{j+1}-2\widetilde{\gamma}_j+\widetilde{\gamma}_{j-1})+A(t_j)\widetilde{\gamma}_j=F_j(I(\{\widetilde{\gamma}_l\}_j)), \quad (15)$$

$F_j(v(\cdot))=(F_j^0(v_j^0(\cdot)),F_j^1(v_j^1(\cdot)),\ldots,F_j^N(v_j^N(\cdot)))^T$, $v(\cdot)=I(\{\widetilde{\gamma}_k\}_j)\in$ Q^{N+1} $[-\tau,\Delta]$; $Q^{N+1}[-\tau,\Delta]$—the space of vector functions with components from $Q[-\tau,\Delta]$.

Operators $A(t_j)$ and $R(t_j)$ are self-adjoint and positive for all t_j [3]. Hence there exists $R^{-1}(t)$. Then we can reduce Eq. (15) to the explicit form

$$\widetilde{\gamma}_{j+1}=2\widetilde{\gamma}_j-\widetilde{\gamma}_{j-1}-R^{-1}(t_j)A(t_j)\widetilde{\gamma}_j+R^{-1}(t_j)(F_j(I(\{\widetilde{\gamma}_l\}_j))) \quad (16)$$

We will assume that the following condition holds

$$R(t)>\frac{1+\epsilon}{4}A(t),t_0\le t\le T, \quad (17)$$

where ϵ is constant independent of Δ and h.

Introduce vector $\gamma_j=(\gamma_j^1,\gamma_j^2)^T=(\widetilde{\gamma}_{j-1},\widetilde{\gamma}_j)^T\in\Gamma$, where Γ is a vector space of dimension $q=2(N+1)$ with norm

$$\|\gamma_n\|_\Gamma=\sqrt{\frac{1}{4}(A(t_{n-1})(\gamma_n^2+\gamma_n^1),\ \gamma_n^2+\gamma_n^1)+((R(t_{n-1})-\frac{1}{4}A(t_{n-1}))(\gamma_n^2-\gamma_n^1),\ \gamma_n^2-\gamma_n^1)} \quad (18)$$

Theorem 2. *For norms (11) and (18) the following inequality holds*

$$\|\gamma_n\|_{\Gamma} \le \frac{D}{\Delta}(\|\gamma_n\|_{\widetilde{\Gamma}} + \|\gamma_{n-1}\|_{\widetilde{\Gamma}}).$$

Proof.

$$\|\gamma_n\|^2_{\Gamma} = \frac{1}{4}(A(t_{n-1})(\gamma_n^2 + \gamma_n^1),\ \gamma_n^2 + \gamma_n^1) + ((R(t_{n-1}) - \frac{1}{4}A(t_{n-1}))(\gamma_n^2 - \gamma_n^1),\ \gamma_n^2 - \gamma_n^1)$$

$$\le\ \frac{1}{4}\|A(t_{n-1})(\gamma_n^2 + \gamma_n^1)\|_{\widetilde{\Gamma}}\|\gamma_n^2 + \gamma_n^1\|_{\widetilde{\Gamma}}$$

$$+\ \|(R(t_{n-1}) - \frac{1}{4}A(t_{n-1}))(\gamma_n^2 - \gamma_n^1)\|_{\widetilde{\Gamma}}\|\gamma_n^2 - \gamma_n^1\|_{\widetilde{\Gamma}}$$

$$\le\ \frac{1}{4\Delta^{\widetilde{\Gamma}}}\|\widetilde{A}(t_{n-1})(\gamma_n^2 + \gamma_n^1)\|_{\widetilde{\Gamma}}\|\gamma_n^2 + \widetilde{\gamma}_n^1\|_{\widetilde{\Gamma}}$$

$$+\ \frac{1}{\Delta^{\widetilde{\Gamma}}}\|(\widetilde{R}(t_{n-1}) - \frac{1}{4}\widetilde{A}(t_{n-1}))(\gamma_n^2 - \gamma_n^1)\|_{\widetilde{\Gamma}}\|\gamma_n^2 - \gamma_n^1\|_{\widetilde{\Gamma}}$$

$$\le\ \frac{1}{4\Delta^2}\|\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}}\|\gamma_n^2 + \gamma_n^1\|^2_{\widetilde{\Gamma}} + \frac{1}{\Delta^2}\|\widetilde{R}(t_{n-1}) - \frac{1}{4}\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}}\|\gamma_n^2 - \gamma_n^1\|^2_{\widetilde{\Gamma}}$$

$$\le\ \frac{1}{4\Delta^2}\|\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}}(\|\gamma_n^2\|_{\widetilde{\Gamma}} + \|\gamma_n^1\|_{\widetilde{\Gamma}})^2$$

$$+\ \frac{1}{\Delta^2}\|\widetilde{R}(t_{n-1}) - \frac{1}{4}\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}}(\|\gamma_n^2\|_{\widetilde{\Gamma}} + \|\gamma_n^1\|_{\widetilde{\Gamma}})^2$$

$$\le\ \left(\frac{1}{4}\|\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}} + \|\widetilde{R}(t_{n-1}) - \frac{1}{4}\widetilde{A}\|_{\widetilde{\Gamma}}\right)\frac{1}{\Delta^2}(\|\gamma_n^2\|_{\widetilde{\Gamma}} + \|\gamma_n^1\|_{\widetilde{\Gamma}})^2.$$

Hence we get the following estimate

$$\|\gamma_n\|_{\Gamma} \le \frac{D}{\Delta}(\|\gamma_n^2\|_{\widetilde{\Gamma}} + \|\gamma_n^1\|_{\widetilde{\Gamma}}),$$

where $D = \sqrt{\frac{1}{4}\|\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}} + \|\widetilde{R}(t_{n-1}) - \frac{1}{4}\widetilde{A}(t_{n-1})\|_{\widetilde{\Gamma}}}$. □

As a result, we obtain the difference scheme:

$$\gamma_{j+1} = S_j\gamma_j + \Delta\Phi(t_j, I(\{\gamma_k\}_j), \Delta), \tag{19}$$

where $S_j = \begin{pmatrix} 0 & 1 \\ -1 & 2 - R^{-1}(t_j)A(t_j) \end{pmatrix}$, $\Phi(t_j, I(\{\gamma_k\}_j), \Delta) = \begin{pmatrix} 0 \\ \frac{R^{-1}(t_j)F_j(I(\{\gamma_k^2\}_j))}{\Delta} \end{pmatrix}$.

The error of approximation (the residual) with interpolation in the general difference scheme is calculated by the formula

$$d_n = (z_{n+1} - S_n z_n)/\Delta - \Phi(t_n, I(\{z_i\}_n)), \; n = 0, ..., M-1. \tag{20}$$

We say that a method has error order $\Delta^{p_1} + h^{p_2}$ for approximation with interpolation if there exists a constant C, such that $\|d_n\|_{\Gamma} \leq C(\Delta^{p_1} + h^{p_2}), n = 0, ..., M-1$.

This definition of residual differs from the earlier introduced definition of residual without interpolation (9). However, the following statement is valid.

Theorem 3. *Suppose that the residual in the sense of (9) has order $\Delta^{p_1} + h^{p_2}$, the functions F^i_j are Lipschitz, interpolation–extrapolation operator I has error order p_0 on the exact solution. Then, the residual with interpolation has the same error order with respect to Δ and h, and this order is $\Delta^{\min\{p_0,p_1,p_2\}}$.*

Proof. Consider the norms of the coordinates of residual (20)

$$\|d^1_n\|^2_{\widetilde{\Gamma}} = \|(z^1_{n+1} - z^2_n)/\Delta\|^2_{\widetilde{\Gamma}} = \left\| \frac{\widetilde{z}_n - \widetilde{z}_n}{\Delta} \right\|^2_{\widetilde{\Gamma}} = 0.$$

$$\begin{aligned}
\|d^2_n\|^2_{\widetilde{\Gamma}} &= \left\| (z^2_{n+1} + z^1_n - 2z^2_n + R^{-1}(t_n)A(t_n)z^2_n)/\Delta - \frac{1}{\Delta} R^{-1}(t_n)(F^n(I(\{z^2_l\}_n))) \right\|^2_{\widetilde{\Gamma}} \\
&= \left\| \frac{\widetilde{z}_{n+1} + \widetilde{z}_{n-1} - 2\widetilde{z}_n}{\Delta} + \frac{1}{\Delta} R^{-1}(t_n)A(t_n)\widetilde{z}_n - \frac{1}{\Delta} R^{-1}(t_n)(F^n(I(\{\widetilde{z}_l\}_n))) \right\|^2_{\widetilde{\Gamma}} \\
&= \Delta^2 \left\| \frac{\widetilde{z}_{n+1} - 2\widetilde{z}_n + \widetilde{z}_{n-1}}{\Delta^2} + \frac{1}{\Delta^2} R^{-1}(t_n)A(t_n)\widetilde{z}_n - \frac{1}{\Delta^2} R^{-1}(t_n)(F^n(I(\{\widetilde{z}_l\}_n)) \right\|^2_{\widetilde{\Gamma}}.
\end{aligned}$$

$$\begin{aligned}
\|d^2_n\|^2_{\widetilde{\Gamma}} &= \Delta^2 \left\| \frac{\widetilde{z}_{n+1} - 2\widetilde{z}_n + \widetilde{z}_{n-1}}{\Delta^2} + \frac{1}{\Delta^2} \widetilde{R}^{-1}(t_n)\widetilde{A}(t_n)\widetilde{z}_n - \widetilde{R}^{-1}(t_n)(F^n(I(\{\widetilde{z}_l\}_n) \right\|^2_{\widetilde{\Gamma}} \\
&\leq \Delta^2 \|\widetilde{R}^{-1}\|^2_{\widetilde{\Gamma}} \left\| \widetilde{R}(t_n) \left(\frac{\widetilde{z}_{n+1} - 2\widetilde{z}_n + \widetilde{z}_{n-1}}{\Delta^2} \right) + \frac{1}{\Delta^2} \widetilde{A}(t_n)\widetilde{z}_n - F^n(I(\{\widetilde{z}_l\}_n)) \right\|^2_{\widetilde{\Gamma}}
\end{aligned} \tag{21}$$

Hence, by the definition of the operators $\widetilde{R}$ (14), $\widetilde{A}$ (13)

$$\left\| \widetilde{R}(t_n)\left(\frac{\widetilde{z}_{n+1}-2\widetilde{z}_n+\widetilde{z}_{n-1}}{\Delta^2}\right)+\frac{1}{\Delta^2}\widetilde{A}(t_n)\widetilde{z}_n - F^n(I(\{\widetilde{z}_l\}_n))\right\|^2_{\widetilde{\Gamma}}$$

$$= \left\| \frac{\widetilde{z}_{n+1}-2\widetilde{z}_n+\widetilde{z}_{n-1}}{\Delta^2} + s\widetilde{A}\left(\frac{\widetilde{z}_{n+1}-2\widetilde{z}_n+\widetilde{z}_{n-1}}{\Delta^2}\right)+\frac{1}{\Delta^2}\widetilde{A}\widetilde{z}_n - F^n(I(\{\widetilde{z}_l\}_n)\right\|^2_{\widetilde{\Gamma}}$$

$$= \left\| \frac{\widetilde{z}_{n+1}-2\widetilde{z}_n+\widetilde{z}_{n-1}}{\Delta^2} + \frac{1}{\Delta^2}(s\widetilde{A}\widetilde{z}_{n+1}+(1-2s)\widetilde{A}\widetilde{z}_n+s\widetilde{A}\widetilde{z}_{n-1}) - F^n(I(\{\widetilde{z}_l\}_n)\right\|^2_{\widetilde{\Gamma}}$$

$$= \sum_{i=1}^{N-1}\Big| \frac{u(x_i,t_{j+1})-2u(x_i,t_j)+u(x_i,t_{j-1})}{\Delta^2} -$$

$$- s\Big(k(x_i+h/2,t_j)\frac{u(x_{i+1},t_{j+1})-u(x_i,t_{j+1})}{h^2} -$$

$$- k(x_i-h/2,t_j)\frac{u(x_i,t_{j+1})-u(x_{i-1},t_{j+1})}{h^2}\Big) +$$

$$- s\Big(k(x_i+h/2,t_j)\frac{u(x_{i+1},t_{j-1})-u(x_i,t_{j-1})}{h^2} -$$

$$- k(x_i-h/2,t_j)\frac{u(x_i,t_{j-1})-u(x_{i-1},t_{j-1})}{h^2}\Big) +$$

$$- (1-2s)\Big(k(x_i+h/2,t_j)\frac{u(x_{i+1},t_j)-u(x_i,t_j)}{h^2} -$$

$$- -k(x_i-h/2,t_j)\frac{u(x_i,t_j)-u(x_{i-1},t_j)}{h^2}\Big) - F^i_n(I(\{u^i_l\}_n))\Big|^2 h. \qquad (22)$$

Let us estimate each term in the sum in (22) using the assumptions of the theorem

$$\Big| \frac{u(x_i,t_{j+1})-2u(x_i,t_j)+u(x_i,t_{j-1})}{\Delta^2} -$$

$$- s\Big(k(x_i+h/2,t_j)\frac{u(x_{i+1},t_{j+1})-u(x_i,t_{j+1})}{h^2} -$$

$$- k(x_i-h/2,t_j)\frac{u(x_i,t_{j+1})-u(x_{i-1},t_{j+1})}{h^2}\Big) +$$

$$- s\Big(k(x_i+h/2,t_j)\frac{u(x_{i+1},t_{j-1})-u(x_i,t_{j-1})}{h^2} -$$

$$- k(x_i-h/2,t_j)\frac{u(x_i,t_{j-1})-u(x_{i-1},t_{j-1})}{h^2}\Big) +$$

$$- (1-2s)\Big(k(x_i+h/2,t_j)\frac{u(x_{i+1},t_j)-u(x_i,t_j)}{h^2} -$$

$$- k(x_i-h/2,t_j)\frac{u(x_i,t_j)-u(x_{i-1},t_j)}{h^2}\Big|$$

$$\leq |\psi^i_n| + |F^i_n(u_{t_n}(x_i,\cdot)) - F^i_n(I(\{u^i_l\}_n))|$$

$$\leq C_1(\Delta^{p_1} + h^{p_2}) + L_F \|u_{t_n}(x_i, \cdot) - I(\{u_l^i\}_n)\|_Q \leq C_1(\Delta^{p_1} + h^{p_2}) + L_F C_2 \Delta^{p_0}. \tag{23}$$

Thus, (21), (22), (23) yield

$$\begin{aligned}\|d_n^2\|_{\widetilde{\Gamma}}^2 &\leq \Delta^2 \|\widetilde{R}^{-1}(t_n)\|_{\widetilde{\Gamma}}^2 \sum_{i=1}^{N-1} (C_1(\Delta^{p_1} + h^{p_2}) + L_F C_2 \Delta^{p_0})^2 h \\ &= \Delta^2 \|\widetilde{R}^{-1}(t_n)\|_{\widetilde{\Gamma}}^2 (N-1)(C_1(\Delta^{p_1} + h^{p_2}) + L_F C_2 \Delta^{p_0})^2 h \\ &\leq \Delta^2 (C_3 \Delta^{\min\{p_1, p_0\}} + C_4 h^{p_2})^2 \\ &= \Delta^2 \left(C_3 \Delta^{\min\{p_1, p_0\}} + C_4 \left(a\Delta/\sqrt{\sigma}\right)^{p_2}\right)^2,\end{aligned}$$

where $C_3 = \|\widetilde{R}^{-1}(t_n)\|_{\widetilde{\Gamma}} \sqrt{X}\,(C_1 + L_F C_2),\ C_4 = \|\widetilde{R}^{-1}(t_n)\|_{\widetilde{\Gamma}} \sqrt{X}\, C_1$.

Therefore

$$\|d_n^2\|_2 \leq C_5 \Delta^{\min\{p_0, p_1, p_2\}+1},$$

where $C_5 = C_3 + C_4(a/\sqrt{\sigma})^{p_2}$.

Using Theorem 2 we obtain

$$\|d_n^2\|_{\widetilde{\Gamma}}^2 \leq \frac{D}{\Delta}(\|d_n\|_{\widetilde{\Gamma}} + \|d_{n-1}\|_{\widetilde{\Gamma}}) \leq \frac{D}{\Delta}(C_5 \Delta^{\min\{p_0, p_1, p_2\}+1}).$$

$$\|d_n\|_{\widetilde{\Gamma}} \leq C \Delta^{\min\{p_0, p_1, p_2\}},$$

where $C = C_5 D$. □

The scheme (19) is stable, if

$$\|S_j\|_\Gamma \leq 1. \tag{24}$$

To investigate the stability of the scheme, we apply the results of [3]. For this, we consider the homogeneous difference scheme corresponding to (15):

$$R(t_j)(\widetilde{\gamma}_{j+1} - 2\widetilde{\gamma}_j + \widetilde{\gamma}_{j-1}) + A(t_j)\widetilde{\gamma}_j = 0 \tag{25}$$

It is proved in [3] that, if the scheme satisfies the conditions (17) and (5), then the solution of (25) satisfies the inequality

$$\|\gamma_{j+1}\|_\Gamma \leq \|\gamma_j\|_\Gamma, \tag{26}$$

which means that the estimate (24) holds.

Using results from [3] and condition (4) we obtain that condition (17) holds if the following condition holds

$$s > \frac{1}{4}\left(1 - \frac{1}{\sigma}\right), \tag{27}$$

where $\sigma = c_2\Delta^2/h^2$.

We have conducted the embedding into the general difference scheme with delay; then we obtain the following statement.

Theorem 4. *Suppose that stability condition (27) holds, the residual in the sense of (9) has order* $\Delta^{p_1} + h^{p_2}$*, the functions* F_j^i *are Lipschitz, the interpolation–extrapolation operator* I *is Lipschitz continuous and has error order* p_0 *on the exact solution. Then, the method converges with order* $\Delta^{\min\{p_0,p_1,p_2\}} + h^{\min\{p_0,p_1,p_2\}}$.

It follows from the theorem that method (8) with the piece-wise linear interpolation (7) converges with order $h^2 + \Delta^2$.

4 Example of Numerical Computation

Consider the equation with delay

$$\frac{\partial^2 u}{\partial t^2}(x,t) = \frac{\partial}{\partial x}\left(\cos xt \cdot \frac{\partial u}{\partial x}\right) + \pi^2 e^{-t}\cos xt \sin \pi x + e^{\tau-2t}\sin^2 \pi x$$

$$+\, \pi t e^{-t}\sin xt \cos(\pi x) + u(x,t)(1 - u(x,t-\tau))\colon 0 \le t \le 3,\ 0 \le x \le 1 \quad (28)$$

for $\tau = 2$, with initial conditions:

$$u(x,t) = e^{-t}\sin \pi x\colon\ -\tau \le t \le 0,\ 0 \le x \le 1$$

and boundary conditions

$$u(x,t) = e^{-t}\sin \pi x : 0 \le t \le 3$$

In Table 1, we present the norms of the differences between the matrices of exact and approximate solutions of Eq. (28) obtained for different values of the parameter s and different steps. The norms of the differences were calculated by the formula

$$\|U\|_1 = \max_{0\le j\le M} \sum_{i=0}^{N} |u(t_j,x_i) - u_j^i|h. \quad (29)$$

Table 1 Norms of differences between the exact and approximate solutions of Eq. (28)

	N = 10	N = 15	N = 20	N = 15	N = 25	N = 50
	M = 60	M = 90	M = 120	M = 36	M = 60	M = 120
s = 0	0.2691	0.1758	24.4750	$9.8 \cdot 10^{11}$	$6.2 \cdot 10^{23}$	$4.7 \cdot 10^{80}$
s = 0.5	0.2280	0.1490	0.1115	0.0013	0.0016	0.0008
s = 1	0.1869	0.1222	0.0615	0.0736	0.0392	0.0192

Acknowledgements This work was supported by RFBR Grant 19-01-00019.

References

1. Lekomtsev, A., Pimenov, V.: Convergence of the scheme with weights for the numerical solution of a heat conduction equation with delay for the case of variable coefficient of heat conductivity. Appl. Math. Comput. **256**, 83–93 (2015)
2. Pimenov, V.G.: General linear methods for numerical solving functional-differential equations. Diff. Equ. **37**(1), 116–127 (2001)
3. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker, New York (2001)
4. Wu, J.: Theory and Applications of Partial Functional Differential Equations. Springer, New York (1996)

Two Nontrivial Solutions for Robin Problems Driven by a p–Laplacian Operator

G. D’Aguì, A. Sciammetta, and E. Tornatore

Abstract By variational methods and critical point theorems, we show the existence of two nontrivial solutions for a nonlinear elliptic problem under Robin condition and when the nonlinearty satisfies the usual Ambrosetti-Rabinowitz condition.

Keywords Robin problem · p-Laplacian · Critical point theory

2010 Mathematics Subject Classification: 35A15 · 35J60

1 Introduction

In this paper we study the existence of two nontrivial weak solutions of following nonlinear elliptic equation under Robin condition

$$\begin{cases} -\Delta_p u + |u|^{p-2}u = \lambda f(x,u) \text{ in } \ \Omega, \\ \frac{\partial u}{\partial \nu} + \beta(x)|u|^{p-2}u = 0 \qquad \text{on } \ \partial\Omega, \end{cases} \tag{1}$$

where $\Omega \subset \mathbf{R}^N$ (with $N \geq 3$) is a non-empty bounded open set with a smooth boundary $\partial\Omega$, λ is a positive real parameter and $1 < p < N$. The differential operator in (1) is described by the p-Laplacian, $\Delta_p u = div(|\nabla u|^{p-2}\nabla u)$. We assume $f : \Omega \times \mathbf{R} \to \mathbf{R}$, $\beta \in L^\infty(\partial\Omega)$, $\beta(x) \geq 0$ a.e. on $\partial\Omega$. In the boundary condition, $\frac{\partial u}{\partial \nu}$ denotes the generalized normal derivative defined by $\frac{\partial u}{\partial \nu} = |\nabla u|^{p-2}\nabla u \cdot \nu(x)$, $\nu(x)$ being the outward unit normal at $x \in \partial\Omega$.

G. D’Aguì
Department of Engineering, University of Messina, 98166 Messina, Italy
e-mail: gdagui@unime.it

A. Sciammetta · E. Tornatore (✉)
Department of Mathematics and Computer Sciences, University of Palermo, 90123 Palermo, Italy
e-mail: elisa.tornatore@unipa.it

A. Sciammetta
e-mail: angela.sciammetta@unipa.it

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_16

A special case of our main result (see Theorem 6) can be given in the following form.

Theorem 1. *Let $g\colon \mathbf{R} \to \mathbf{R}$ be a nonnegative and continuous function such that there exist positive constants a_1, a_2 and $s \in]p, p^*[$ such that*

$$|g(t)| \leq a_1 + a_2|t|^{s-1} \quad \textit{for all } t \in \mathbf{R},$$

and

$$\lim_{\tau \to 0^+} \frac{g(\tau)}{\tau} = +\infty.$$

Moreover, assume that there exist $\nu > p$ and $R > 0$ such that

$$0 < \nu \int_0^\tau g(t)dt \leq \tau g(\tau) \quad \textit{for all } \tau \in \mathbf{R} \textit{ with } |\tau| \geq R.$$

Then, there exists $\overline{\lambda} > 0$ such that for each $\lambda \in]0, \overline{\lambda}[$, the problem

$$\begin{cases} -\Delta_p u + |u|^{p-2}u = \lambda g(u) \ \textit{in} \ \ \Omega, \\ \frac{\partial u}{\partial \nu} + \beta(x)|u|^{p-2}u = 0 \qquad \textit{on} \ \ \partial\Omega, \end{cases} \tag{2}$$

has at least two nonnegative weak solutions.

The main novelty of our paper is that we apply a recent critical-points result to elliptic problems with p–Laplacian in the equation and with Robin conditions on the boundary. There exist several existence results to problem (1), anyway our approach is new and gives the existence of two nontrivial weak solutions. The assumptions on the nonlinear term are easy to verify and so our results could be applied to several problems of type (1).

Elliptic problems with Robin conditions have been studied by several authors by applying different tools like fixed point theorems, sub and super-solution methods, and critical point theory. We refer, without any claim to completeness, to the papers [2, 7, 12–15] and the references therein.

Moreover, we observe that the derivation and application of critical point results of that used here have been initiated by the works of Ricceri [16, 17] which were the starting point of several generalizations in that direction for smooth and non-smooth functionals, we refer only to some works of Marano-Motreanu [9, 10], and Bonanno [3, 4] that inspired us in writing this paper.

The paper is organized as follows. In Sect. 2, we state the main definitions and tools that we are going to need to prove our main results. Especially, we recall the abstract critical point theorem of Bonanno-D'Aguì [5], which is an appropriate combination of the local minimum theorem obtained by Bonanno with the classical and seminal Ambrosetti–Rabinowitz theorem (see [1]), moreover we give a lemma about the relation of our perturbation concerning the Ambrosetti–Rabinowitz condition and the Palais-Smale condition (Lemma 1). Then, in Sect. 3, we are going to prove our main result which gives an answer about the existence of solutions to problem (1). To be more precise, we obtain the existence of two nontrivial solutions of (1), see Theorem 3, and the proof is based on the abstract critical points result stated in Sect. 2. Finally, in Sect. 4, we consider special problem in the autonomous case, and give an example in order to show the applicability of our results.

2 Preliminaries and Basic Notations

Let $(X, \|\cdot\|)$ be a Banach space; its dual space is X^* and the corresponding duality pairing is denoted by $\langle\cdot,\cdot\rangle$. Let $I : X \to \mathbf{R}$ be a Gâteaux differentiable functional; we say that I satisfies the Palais-Smale condition, (in short (PS)–condition), if every sequence $\{u_n\}_{n\in\mathbf{N}} \subseteq X$ such that $\{I(u_n)\}_{n\in\mathbf{N}} \subset \mathbf{R}$ is bounded, and $I'(u_n) \to 0$ in X^* as $n \to +\infty$, admits a strongly convergent subsequence in X.

Let $A : X \to X^*$ be a functional. We say that A has S_+-property iff every sequence $\{u_n\}_{n\in\mathbf{N}} \subset X$ such that $u_n \rightharpoonup u$ in X and $\limsup_{n\to+\infty}\langle Au_n, u_n - u\rangle \le 0$ implies that $u_n \to u$ in X.

We consider the usual Sobolev space $W^{1,p}(\Omega)$, endowed with the norm

$$\|u\| = \left(\int_\Omega |u(x)|^p dx + \int_\Omega |\nabla u(x)|^p dx\right)^{1/p},$$

and denote by $(W^{1,p}(\Omega))^*$ its dual space.

Since $1 < p < N$, $p^* = \frac{pN}{N-p}$ and it is known that, for every $u \in W^{1,p}(\Omega)$ there exists a constant $T \in \mathbf{R}_+$ such that

$$\|u\|_{L^{p^*}(\Omega)} \le T\|u\|, \tag{3}$$

the constat T has been determined by Talenti (see [18]) and

$$T \le \pi^{-\frac{1}{2}} N^{-\frac{1}{p}} \left(\frac{p-1}{N-p}\right)^{1-\frac{1}{p}} \left(\frac{\Gamma\left(1+\frac{N}{2}\right)\Gamma(N)}{\Gamma\left(\frac{N}{p}\right)\Gamma\left(1+N-\frac{N}{p}\right)}\right)^{\frac{1}{N}},$$

where Γ is the Euler function.

Fix $s \in [1, p^*[$, by Sobolev embedding theorem and Hölder's inequality, for every $u \in W^{1,p}(\Omega)$ we have that

$$\|u\|_{L^s(\Omega)} \le T|\Omega|^{\frac{p^*-s}{p^*s}}\|u\|, \tag{4}$$

where $|\Omega|$ denotes the Lebesgue measure of Ω in $\mathbf{R}$. On $\partial\Omega$ we consider the $(N-1)$-dimensional Hausdorff (surface) measure $\sigma(\cdot)$. Using this measure, we can define in the usual way the "boundary" Lebesgue spaces $L^p(\partial\Omega)$ $1 \le p \le \infty$. From the theory of Sobolev spaces, we know that there exists a unique continuous linear map $\gamma_0 : W^{1,p}(\Omega) \to L^p(\partial\Omega)$, known as the "trace map", such that

$$\gamma_0(u) = u_{|\partial\Omega} \text{ for all } u \in W^{1,p}(\Omega) \cap C(\overline{\Omega}).$$

Therefore we understand $\gamma_0(u)$ as representing the "boundary values" of an arbitrary Sobolev function u. The trace map γ_0 is compact into $L^\eta(\partial\Omega)$ for all $\eta \in \left[1, \frac{(N-1)p}{N-p}\right[$. Also, we have

$$\mathrm{im}\gamma_0 = W^{\frac{1}{p'},p}(\partial\Omega), \; \left(p' = \frac{p}{p-1}\right), \quad \ker\gamma_0 = W^{1,p}(\Omega).$$

In the sequel, for the sake of notational simplicity, we drop the use of the trace map γ_0. All restrictions of Sobolev functions u on $\partial\Omega$ are defined in the sense of traces. In studying problem (1) we rely on the negative p-Laplacian $-\Delta_p : W^{1,p}(\Omega) \to (W^{1,p}(\Omega))^*$. It is well-known that the operator $-\Delta_p$ is continuous, bounded, pseudomonotone and has the S_+-property (see [6, 11]).

Throughout the sequel, we assume that the nonlinearity $f : \Omega \times \mathbf{R} \to \mathbf{R}$ is a Carathéodory function i.e. $f(\cdot, t)$ is measurable for every $t \in \mathbf{R}$, $f(x, \cdot)$ is continuous for almost every $x \in \Omega$ and satisfies the subcritical growth condition and the usual Ambrosetti-Rabinowitz condition (in short (AR)-condition).

(H) There exist two nonnegative constants a_1, a_2, a constant $s \in]p, p^*[$ such that

$$|f(x,t)| \le a_1 + a_2|t|^{s-1} \quad \text{for all } (x,t) \in \Omega \times \mathbf{R}.$$

Put $F(x,t) = \displaystyle\int_0^t f(x,\xi)d\xi$ for all $(x,t) \in \Omega \times \mathbf{R}$.

(AR) There exist two constants $\mu > p$ and $M > 0$ such that, $0 < \mu F(x,t) \le tf(x,t)$, for all $x \in \Omega$ and for all $|t| \ge M$.

We consider the C^1-functionals $\Phi,\ \Psi : W^{1,p}(\Omega) \to \mathbf{R}$ defined by

$$\Phi(u) = \frac{1}{p}\|u\|^p + \frac{1}{p}\int_{\partial\Omega} \beta(x)|u(x)|^p d\sigma, \tag{5}$$

and

$$\Psi(u) = \int_\Omega F(x, u(x))dx, \tag{6}$$

for all $u \in W^{1,p}(\Omega)$, whose Gâteaux derivatives at point $u \in W^{1,p}(\Omega)$ are given by

$$\Phi'(u)(v) = \int_{\Omega} |\nabla u(x)|^{p-2}\nabla u(x) \cdot \nabla v(x)dx$$

$$+ \int_{\Omega} |u(x)|^{p-2}u(x)v(x)dx + \int_{\partial\Omega} \beta(x)|u(x)|^{p-2}uvd\sigma,$$

and

$$\Psi'(u)(v) = \int_{\Omega} f(x, u(x))v(x)dx,$$

for every $v \in W^{1,p}(\Omega)$. Put $I_\lambda = \Phi - \lambda\Psi$, we observe that critical points of I_λ are weak solutions of (1).

We recall that a weak solution of problem (1) is any $u \in W^{1,p}(\Omega)$ such that

$$\int_{\Omega} |\nabla u(x)|^{p-2}\nabla u(x) \cdot \nabla v(x)dx + \int_{\Omega} |u(x)|^{p-2}u(x)v(x)dx$$

$$+ \int_{\partial\Omega} \beta(x)|u(x)|^{p-2}u(x)v(x)d\sigma = \lambda \int_{\Omega} f(x, u(x))v(x)dx.$$

Finally, we recall the following two non-zero critical points theorem established in [5] that we use to point out our results.

Theorem 2. *Let X be a real Banach space and let $\Phi, \Psi : X \to \mathbf{R}$ be two functionals of class C^1 such that $\inf_X \Phi(u) = \Phi(0) = \Psi(0) = 0$. Assume that there are $r \in \mathbf{R}$ and $\tilde{u} \in X$, with $0 < \Phi(\tilde{u}) < r$, such that*

$$\frac{\sup\limits_{u\in\Phi^{-1}(]-\infty,r])} \Psi(u)}{r} < \frac{\Psi(\tilde{u})}{\Phi(\tilde{u})}, \tag{7}$$

and, for each

$$\lambda \in \Lambda = \left]\frac{\Phi(\tilde{u})}{\Psi(\tilde{u})}, \frac{r}{\sup\limits_{u\in\Phi^{-1}(]-\infty,r])} \Psi(u)}\right[,$$

the functional $I_\lambda = \Phi - \lambda\Psi$ satisfies the (PS)–condition and it is unbounded from below.

Then, for each $\lambda \in \Lambda$, the functional I_λ admits at least two non-zero critical points $u_{\lambda,1}, u_{\lambda,2} \in X$ such that $I(u_{\lambda,1}) < 0 < I(u_{\lambda,2})$.

3 Main Results

In this section, we present our main results. To be precise, we establish the existence result of two non zero weak solutions of problem (1).

We have the following Lemma.

Lemma 1. *Assume that conditions (H)-(AR) hold. Then I_λ satisfies the (PS)–condition.*

Proof. Let $\{u_n\}_{n\in\mathbf{N}} \subseteq W^{1,p}(\Omega)$ be a sequence such that $\{I_\lambda(u_n)\}_{n\in\mathbf{N}} \subset \mathbf{R}$ is bounded, and $I'_\lambda(u_n) \to 0$ in $(W^{1,p}(\Omega))^*$ as $n \to +\infty$. Simple calculations show that

$$\begin{aligned}\mu I_\lambda(u_n) - ||I'_\lambda(u_n)||_{(W^{1,p}(\Omega))^*}||u_n|| &\geq \mu I_\lambda(u_n) - I'_\lambda(u_n)(u_n) \\ &= \mu\Phi(u_n) - \lambda\mu\Psi(u_n) - \Phi'(u_n)(u_n) + \lambda\Psi'(u_n)(u_n) \\ &= \left(\frac{\mu}{p}-1\right)||u_n||^p + \left(\frac{\mu}{p}-1\right)\int_{\partial\Omega}\beta(x)|u_n(x)|^p d\sigma \\ &- \lambda\int_\Omega (\mu F(x,u_n(x)) - f(x,u_n(x))u_n(x))\,dx \\ &\geq \left(\frac{\mu}{p}-1\right)||u_n||^p + C,\end{aligned} \tag{8}$$

where C is a constant. If $\{u_n\}_{n\in\mathbf{N}}$ is not bounded, from (8) we obtain a contradiction. Therefore $\{u_n\}_{n\in\mathbf{N}}$ is bounded in $W^{1,p}(\Omega)$. Then, using a subsequence if necessary we may assume that $u_n \rightharpoonup u$ in $W^{1,p}(\Omega)$, $u_n \to u$ in $L^l(\Omega)$ where $l \in [1, p^*[$ and $u_n \to u$ in $L^\eta(\partial\Omega)$ for $\eta \in \left[1, \frac{(N-1)p}{N-p}\right[$.

Using (H) and the Hölder inequality, we obtain that

$$\lim_{n\to\infty}\int_\Omega f(x,u_n)(u_n-u)dx = 0, \tag{9}$$

$$\lim_{n\to\infty}\int_{\partial\Omega}\beta(x)|u_n|^{p-2}u_n(u_n-u)d\sigma = 0, \tag{10}$$

and

$$\lim_{n\to\infty}\int_\Omega |u_n|^{p-2}u_n(u_n-u)dx = 0. \tag{11}$$

Taking into account that such that $I'_\lambda(u_n) \to 0$ in X^* as $n \to +\infty$, we have that

$$\begin{aligned}\langle I'_\lambda(u_n), u_n - u\rangle &= \langle -\Delta_p u_n, u_n - u\rangle + \int_\Omega |u_n|^{p-2}u_n(u_n-u)dx \\ &+ \int_{\partial\Omega}\beta(x)|u_n|^{p-2}u_n(u_n-u)d\sigma - \int_\Omega f(x,u_n)(u_n-u)dx \to 0.\end{aligned}$$

From (9), (10) and (11) one has

$$\limsup_{n\to\infty}\langle -\Delta_p u_n, u_n - u\rangle \leq 0.$$

By the S_+-property of $-\Delta_p$ in $W^{1,p}(\Omega)$ we have that $u_n \to u$ in $W^{1,p}(\Omega)$. Hence I_λ fulfills (PS)–condition. □

Put

$$k = \frac{|\Omega| + \beta_\infty|\partial\Omega|}{|\Omega|^{\frac{p}{p^*}}}T^p, \tag{12}$$

where $|\partial\Omega| = \displaystyle\int_{\partial\Omega} d\sigma = \sigma(\partial\Omega)$ and $\beta_\infty = ess\sup_{\Omega}\beta(x)$.

Theorem 3. *Assume that conditions (H) and (AR) hold. Moreover assume that there are two positive constants c and d, with $d < c$, such that*

$$a_1c^{1-p} + \frac{a_2}{s}c^{s-p} < \frac{1}{k|\Omega|}\frac{\displaystyle\int_\Omega F(x,d)dx}{d^p}, \tag{13}$$

where a_1, a_2, s and k are given by (H) and (12) respectively.

Then, for each $\lambda \in \Lambda_1 := \left]\frac{k|\Omega|^{\frac{p}{p^*}}}{pT^p}\frac{d^p}{\displaystyle\int_\Omega F(x,d)dx}, \frac{1}{pT^p|\Omega|^{\frac{p}{N}}}\frac{1}{a_1c^{1-p}+\frac{a_2}{s}c^{s-p}}\right[$,
problem (1) has at least two non-zero weak solutions.

Proof. Put Φ and Ψ as in (5) and (6). It is well known that Φ and Ψ satisfy all regularity assumptions requested in Theorem 2.

Explicitly, we observe that from (13), one has $\Lambda_1 \neq \emptyset$.

Consider the constant function $\overline{u}(x) = dforallx \in \Omega$, we observe that $\overline{u} \in W^{1,p}(\Omega)$, taking into account (12) we have

$$\Phi(\overline{u}) = \frac{d^p}{p}\left(\int_\Omega dx + \int_{\partial\Omega}\beta(x)d\sigma\right) \leq \frac{d^p}{p}\left(|\Omega| + \beta_\infty|\partial\Omega|\right) = \frac{k|\Omega|^{\frac{p}{p^*}}}{pT^p}d^p. \tag{14}$$

On the other hand one has

$$\Psi(\overline{u}) = \int_\Omega F\,(x,d)\,dx,$$

hence, we obtain

$$\frac{\Psi(\overline{u})}{\Phi(\overline{u})} \geq \frac{pT^p}{k|\Omega|^{\frac{p}{p^*}}}\frac{\displaystyle\int_\Omega F\,(x,d)\,dx}{d^p}. \tag{15}$$

Now, set $r = \frac{1}{p}\frac{|\Omega|^{\frac{p}{p^*}}}{T^p}c^p$. For all $u \in W^{1,p}(\Omega)$ such that $u \in \Phi^{-1}\left(]-\infty, r]\right)$, taking (5) into account, one has that $\|u\| \leq (pr)^{\frac{1}{p}}$ we have

$$\Phi^{-1}\left(]-\infty, r]\right) \subseteq \left\{u \in W^{1,p}(\Omega) : \|u\| \leq (pr)^{\frac{1}{p}}\right\}. \tag{16}$$

From (H) follows

$$|F(x,t)| \leq a_1|t| + a_2\frac{|t|^s}{s} \text{ for every } (x,t) \in \Omega \times \mathbf{R}. \tag{17}$$

From (4), (16) and (17) one has

$$\begin{aligned}
\frac{\sup\limits_{u \in \Phi^{-1}(]-\infty,r])} \Psi(u)}{r} &\leq \frac{\sup\limits_{\|u\| \leq (pr)^{\frac{1}{p}}} \Psi(u)}{r} \qquad (18)\\
&\leq \frac{\sup\limits_{\|u\| \leq (pr)^{\frac{1}{p}}} \left(a_1\|u\|_{L^1(\Omega)} + \frac{a_2}{s}\|u\|^s_{L^s(\Omega)}\right)}{r}\\
&\leq \frac{\sup\limits_{\|u\| \leq (pr)^{\frac{1}{p}}} \left(a_1 T|\Omega|^{\frac{p^*-1}{p^*}}\|u\| + \frac{a_2}{s}T^s|\Omega|^{\frac{p^*-s}{p^*}}\|u\|^s\right)}{r}\\
&\leq \frac{a_1 T|\Omega|^{\frac{p^*-1}{p^*}}(pr)^{\frac{1}{p}} + \frac{a_2}{s}T^s|\Omega|^{\frac{p^*-s}{p^*}}(pr)^{\frac{s}{p}}}{r}\\
&= pT^p|\Omega|^{\frac{p^*-p}{p^*}}\left[a_1\left(\frac{T^p pr}{|\Omega|^{\frac{p}{p^*}}}\right)^{\frac{1-p}{p}} + \frac{a_2}{s}\left(\frac{T^p pr}{|\Omega|^{\frac{p}{p^*}}}\right)^{\frac{s-p}{p}}\right]\\
&= pT^p|\Omega|^{\frac{p}{N}}\left[a_1c^{1-p} + \frac{a_2}{s}c^{s-p}\right].
\end{aligned}$$

Therefore, from (13), (15), (18) we obtain condition (7) of Theorem 2. Moreover, since $0 < d < c$ and again by virtue of (13), we infer that

$$kd^p < c^p. \tag{19}$$

Indeed, arguing by contradiction, if we assume that $kd^p \geq c^p$ and using (17) we have

$$a_1c^{1-p} + \frac{a_2}{s}c^{s-p} \geq \frac{1}{k}\frac{a_1d + \frac{a_2}{s}d^s}{d^p} \geq \frac{1}{k|\Omega|}\frac{\int_\Omega F(x,d)\,dx}{d^p},$$

which contradicts (13). Then from (14), (19) we obtain that

$$\Phi(\overline{u}) < r.$$

By virtue of Lemma 1, for all fix $\lambda \in \Lambda_1$ the functional I_λ satisfies the (PS)–condition. Using (AR)–condition, it is easy to prove that the functional I_λ is unbounded from below. Moreover, $\inf_{u \in W^{1,p}(\Omega)} \Phi(u) = \Phi(0) = \Psi(0) = 0$, therefore, all assumptions of Theorem 2 are satisfied. So, for all $\lambda \in \Lambda_1 \subset \Lambda$ problem (1) admits at least two non-zero weak solutions. □

Finally, we point out the following result that we will use to obtain nonnegative solutions for our problem (1).

Lemma 2. *Let $f : \Omega \times \mathbf{R} \to \mathbf{R}$, assume that $f(x, 0) \geq 0$ for a.e. $x \in \Omega$. Consider the problem*

$$\begin{cases} -\Delta_p u + |u|^{p-2}u = \lambda f_+(x, u) \ in \ \Omega, \\ \frac{\partial u}{\partial \nu} + \beta(x)|u|^{p-2}u = 0 \qquad\qquad on \ \partial\Omega, \end{cases} \tag{20}$$

where

$$f_+(x, t) = \begin{cases} f(x, 0), \ if \ t < 0, \\ f(x, t), \ if \ t \geq 0. \end{cases} \tag{21}$$

Then, the weak solutions of problem (20) are nonnegative weak solution of problem (1).

Proof. If $\bar{u} \in W^{1,p}(\Omega)$ is a weak solution of (20), choosing $v = \bar{u}^- = \max\{-u, 0\} \in W^{1,p}(\Omega)$ as test function (see, for instance, [8, Lemma 7.6]), one has

$$\int_{\{\bar{u}<0\}} |\nabla \bar{u}(x)|^p dx + \int_{\{\bar{u}<0\}} |\bar{u}(x)|^p dx + \int_{\partial\Omega} \beta(x)|\bar{u}(x)|^p d\sigma$$

$$= \lambda \int_{\{\bar{u}<0\}} f_+(x, \bar{u}(x))\bar{u}(x)dx \leq 0,$$

that is $\bar{u} \geq 0$ for a.e. $x \in \Omega$. Then $\bar{u}$ is a nonnegative weak solution of problem (1) Hence, our claim is proved. □

Now, we present our result on the existence of at least two nonnegative solutions.

Theorem 4. *Let $f : \Omega \times \mathbf{R} \to \mathbf{R}$ be a continuous functions, $f(x, 0) \geq 0$ a. e. $x \in \Omega$. Assume that (H) and (AR)–condition hold. Moreover, there are two positive constants c and d, with $d < c$, such that*

$$a_1 c^{1-p} + \frac{a_2}{s} c^{s-p} < \frac{1}{k|\Omega|} \frac{\int_\Omega F(x, d)dx}{d^p}. \tag{22}$$

Then, for each $\lambda \in \Lambda_1 := \left]\frac{k|\Omega|^{\frac{p}{p^*}}}{pT^p}\frac{d^p}{\displaystyle\int_\Omega F(x,d)dx}, \frac{1}{pT^p|\Omega|^{\frac{p}{N}}}\frac{1}{a_1c^{1-p}+\frac{a_2}{s}c^{s-p}}\right[$ *problem (1) has at least two nontrivial and nonnegative solutions.*

Proof. Since all conditions of Theorem 3 are satisfied, then for each $\lambda \in \Lambda_1$ the problem (1) admits at least two non zero weak solutions in $W^{1,p}(\Omega)$ and, taking into account Lemma 2, they are also nonnegative. □

4 Some Consequences

We point out a special case of Theorem 3 when the nonlinearity f does not depend on x.

Theorem 5. *Let* $f : \mathbf{R} \to \mathbf{R}$ *be a nonnegative continuous function such that* (H) *and* (AR)*–condition hold. Moreover, assume that there are two positive constants c and d, with* $d < c$*, such that*

$$a_1c^{1-p} + \frac{a_2}{s}c^{s-p} < \frac{1}{k}\frac{F(d)}{d^p}. \tag{23}$$

Then, for each $\lambda \in \Lambda_2 := \left]\frac{k}{pT^p|\Omega|^{\frac{p}{N}}}\frac{d^p}{F(d)}, \frac{1}{pT^p|\Omega|^{\frac{p}{N}}}\frac{1}{a_1c^{1-p}+\frac{a_2}{s}c^{s-p}}\right[$ *problem*

$$\begin{cases} -\Delta_p u + |u|^{p-2}u = \lambda f(u) \ in \ \Omega, \\ \frac{\partial u}{\partial \nu} + \beta(x)|u|^{p-2}u = 0 \qquad on \ \partial\Omega, \end{cases} \tag{24}$$

has at least two nonnegative weak solutions.

Proof. Our aim is to apply Theorem 4. We observe that from condition (23) we obtain condition (13) of Theorem 3 and moreover $f(x,0) \geq 0$ a.e. $x \in \Omega$. Then, for each $\lambda \in \Lambda_2 := \left]\frac{k}{pT^p|\Omega|^{\frac{p}{N}}}\frac{d^p}{F(d)}, \frac{1}{pT^p|\Omega|^{\frac{p}{N}}}\frac{1}{a_1c^{1-p}+\frac{a_2}{s}c^{s-p}}\right[$ problem (24) has at least two nonnegative weak solutions. □

Finally, we want to consider the case when the nonlinear term of problem (24) is super-$(p-1)$ linear at zero.

Theorem 6. *Let* $f : \mathbf{R} \to \mathbf{R}$ *be a nonnegative continuous function such that* (H) *and* (AR)*–condition hold and*

$$\limsup_{t\to 0^+} \frac{F(t)}{t^p} = +\infty, \tag{25}$$

and put $\lambda^* = \frac{1}{pT^p|\Omega|^{\frac{p}{N}}} \sup_{c>0} \frac{1}{a_1 c^{1-p} + \frac{a_2}{s} c^{s-p}}$.

Then, for each $\lambda \in]0, \lambda^*[$*, problem (24) admits at least two nonnegative weak solutions.*

Proof. Put $\lambda \in]0, \lambda^*[$, there is $c > 0$ such that $\lambda < \frac{1}{pT^p|\Omega|^{\frac{p}{N}}} \frac{1}{a_1 c^{1-p} + \frac{a_2}{s} c^{s-p}}$. From (25) there is $0 < d < c$ such that $\frac{pT^p|\Omega|^{\frac{p}{N}}}{k} \frac{F(d)}{d^p} > \frac{1}{\lambda}$. Hence, Theorem 5 guarantees the conclusion. □

Example 1. Let $p = 3$, $N = 4$ and $\Omega = B(0, 3^{\frac{1}{8}})$, the open ball of radius $r = 3^{\frac{1}{8}}$ and consider the function $f : \mathbf{R} \to \mathbf{R}$ given by $f(t) = t^4 + 1$.

Putting $a_1 = 1$, $a_2 = 5$ and $s = 5$, we observe that conditions (H) holds. On the other hand

$$F(t) = \int_0^t (\xi^4 + 1) d\xi = \frac{t^5}{5} + t\,,$$

$$\limsup_{t \to 0^+} \frac{F(t)}{t^p} = \lim_{t \to 0^+} \frac{t^5 + 5t}{5t^3} = +\infty,$$

and (AR)–condition is satisfied as a simple computation shows.
Moreover, one has that

$$T \leq \pi^{-\frac{1}{2}} 4^{-\frac{1}{3}} 2^{\frac{2}{3}} \left(\frac{\Gamma(3)\Gamma(4)}{\Gamma\left(\frac{4}{3}\right)\Gamma\left(\frac{11}{3}\right)} \right)^{\frac{1}{4}},$$

$$\sup_{c>0} \frac{1}{a_1 c^{1-p} + \frac{a_2}{s} c^{s-p}} = \sup_{c>0} \frac{1}{\frac{1}{c^2} + c^2} = \frac{1}{2},$$

$$\lambda^* = \frac{1}{pT^p|\Omega|^{\frac{p}{N}}} \sup_{c>0} \frac{1}{a_1 c^{1-p} + \frac{a_2}{s} c^{s-p}} \geq \frac{2^2 \cdot 5^{\frac{3}{4}} \cdot \pi^{\frac{3}{4}}}{3^{\frac{11}{2}}}.$$

Using Theorem 6, for each $\lambda \in \left]0, \frac{2^2 \cdot 5^{\frac{3}{4}} \cdot \pi^{\frac{3}{4}}}{3^{\frac{11}{2}}}\right[$, the problem

$$\begin{cases} -\Delta_3 u + |u|u = \lambda(t^4 + 1) & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} + \beta(x)|u|u = 0 & \text{on } \partial\Omega, \end{cases}$$

admits at least two nonnegative weak solutions.

Acknowledgements The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The paper is partially supported by PRIN 2017—Progetti di Ricerca di rilevante Interesse Nazionale, "Nonlinear Differential Problems via Variational, Topological and Set-valued Methods" (2017AYM8XW).

References

1. Ambrosetti, A., Rabinowitz, P.H.: Dual variational methods in critical point theory and applications. J. Funct. Anal. **14**, 349–381 (1973)
2. Averna, D., Papageorgiou, N.S., Tornatore, E.: Positive solutions for nonlinear Robin problems. Electron. J. Differ. Equ. **2017**(204), 1–25 (2017)
3. Bonanno, G.: Relations between the mountain pass theorem and local minima. Adv. Nonlinear Anal. **1**, 205–220 (2012)
4. Bonanno, G.: A critical point theorem via the Ekeland variational principle. Nonlinear Anal. **75**(5), 2992–3007 (2012)
5. Bonanno, G., D'Aguì, G.: Two non-zero solutions for elliptic Dirichlet problems. Z. Anal. Anwend **35**, 449–464 (2016)
6. Carl, S., Le, V.K., Motreanu, D.: Nonsmooth Variational Problems and Their Inequalities. Comparison Principles and Applications. Springer, New York (2007)
7. Drabek, P., Schindler, I.: Positive solutions for the p-Laplacian with Robin boundary conditions on irregular domains. Appl. Math. Lett. **24**, 588–591 (2011)
8. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, 2nd edn. Springer, Berlin (1983)
9. Marano, S.A., Motreanu, D.: Infinitely many critical points of non-differentiable functions and applications to a Neumann type problem involving the p-Laplacian. J. Differ. Equ. **182**, 108–120 (2002)
10. Marano, S.A., Motreanu, D.: On a three critical points theorem for non-differentiable functions and applications to nonlinear boundary value problems. Nonlinear Anal. **48**(1), 37–52 (2002)
11. Motreanu, D., Motreanu, V.V., Papageorgiou, N.S.: Topological and Variational Methods with Applications to Nonlinear Boundary Value Problems. Springer, New York (2014)
12. Motreanu, D., Sciammetta. A., Tornatore, E.: A sub-supersolution approach for Robin boundary value problem with full gradient dependence. Mathematics **658**(8) (2020)
13. Papageorgiou, N.S., Radulescu, V.D.: Multiple solutions with precise sign for nonlinear parametric Robin problems. J. Differ. Equ. **256**, 2449–2479 (2014)
14. Papageorgiou, N.S., Radulescu, V.D.: Bifurcation near infinity for the Robin p-Laplacian. Manuscripta Math. **148**, 415–433 (2015)
15. Papageorgiou, N.S., Radulescu, V.D., Repovs, D.: Positive solutions for nonlinear nonhomogeneous Robin problems. Forum Math. **30**, 553–580 (2018)
16. Ricceri, B.: A general variational principle and some of its applications. J. Comput. Appl. Math. **113**(1–2), 401–410 (2000)
17. Ricceri, B.: On a three critical points theorem. Arch. Math. (Basel) **75**(3), 220–226 (2000)
18. Talenti, G.: Best constant in Sobolev inequality. Ann. Mat. Pura Appl. **110**, 353–372 (1976)

Multiple Periodic Solutions for a Duffing Type Equation with One-Sided Sublinear Nonlinearity: Beyond the Poincaré-Birkhoff Twist Theorem

Tobia Dondè and Fabio Zanolin

Abstract We prove the existence of multiple periodic solutions for a planar Hamiltonian system generated from the second order scalar ODE of Duffing type $x'' + q(t)g(x) = 0$ with g satisfying a one-sided condition of sublinear type. We consider the classical approach based on the Poincaré-Birkhoff fixed point theorem as well as some refinements on the side of the theory of bend-twist maps and topological horseshoes. We focus our analysis to the case of a stepwise weight function, in order to highlight the underlying geometrical structure.

1 Introduction

The Poincaré-Birkhoff fixed point theorem deals with a planar homeomorphism Ψ defined on an annular region A, such that Ψ is area-preserving, leaves the boundary of A invariant and rotates the two components of ∂A in opposite directions (twist condition). Under these assumptions, in 1912 Poincaré conjectured (and proved in some particular cases) the existence of at least two fixed points for Ψ, a result known as "the Poincaré last geometric theorem". A proof for the existence of at least one fixed point (and actually two in a non-degenerate situation) was obtained by Birkhoff in 1913 [3]. In the subsequent years Birkhoff reconsidered the theorem as well as its possible extensions to a more general setting, for instance, removing the assumption of boundary invariance, or proposing some hypotheses of topological nature instead of the area-preserving condition, thus opening a line of research that is still active today (see for example [4, 9], and the references therein). The skepticism of some mathematicians about the correctness of the proof of the second fixed point motivated Brown and Neumann to present in [7] a full detailed proof, adapted from Birkhoff's 1913 paper, in order to eliminate previous possible controversial aspects. Another approach for the proof of the second fixed point has been proposed in [47], coupling [3] with a result for removing fixed points of zero index.

T. Dondè · F. Zanolin (✉)
Università di Udine, Via delle Scienze 206, 33100 Udine, Italy
e-mail: fabio.zanolin@uniud.it

T. Dondè
e-mail: donde.tobia@spes.uniud.it

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_17

In order to express the twist condition in a more precise manner, the statement of the Poincaré-Birkhoff theorem is usually presented in terms of the lifted map $\tilde{\Psi}$. Let us first introduce some notation. Let $D(R)$ and $D[R]$ be, respectively, the open and the closed disc of center the origin and radius $R > 0$ in $\mathbb{R}^2$ endowed with the Euclidean norm $||\cdot||$. Let also $C_R := \partial D(R)$. Given $0 < r < R$, we denote by A or $A[r, R]$ the closed annulus $A[r, R] := D[R] \setminus D(r)$. Hence the area-preserving (and orientation-preserving) homeomorphism $\Psi : A \to \Psi(A) = A$ is lifted to a map $\tilde{\Psi} : \tilde{A} \to \tilde{A}$, where $\tilde{A} := \mathbb{R} \times [r, R]$ is the covering space of A via the covering projection $\Pi : (\theta, \rho) \mapsto (\rho\cos\theta, \rho\sin\theta)$ and

$$\tilde{\Psi} : (\theta, \rho) \mapsto (\theta + 2\pi\mathscr{J}(\theta, \rho), \mathscr{R}(\theta, \rho)), \tag{1}$$

with the functions $\mathscr{J}$ and $\mathscr{R}$ being 2π-periodic in the θ-variable. Then, the classical (1912–1913) Poincaré-Birkhoff fixed point theorem can be stated as follows (see [7]).

Theorem 1. *Let $\Psi : A \to \Psi(A) = A$ be an area preserving homeomorphism such that the following two conditions are satisfied:*

$$\mathscr{R}(\theta, r) = r, \quad \mathscr{R}(\theta, R) = R, \quad \forall\,\theta \in \mathbb{R}, \tag{$PB1$}$$

$$\exists\, j \in \mathbb{Z} : \; (\mathscr{J}(\theta, r) - j)(\mathscr{J}(\theta, R) - j) < 0, \quad \forall\,\theta \in \mathbb{R}. \tag{$PB2$}$$

Then Ψ has at least two fixed points z_1, z_2 in the interior of A and $\mathscr{J}(\theta, \rho) = j$ for $\Pi(\theta, \rho) = z_i$.

We refer to condition $(PB1)$ as to the “boundary invariance” and we call $(PB2)$ the “twist condition”. The function $\mathscr{J}$ can be regarded as a rotation number associated with the points. In the original formulation of the theorem it is $j = 0$, however any integer j can be considered.

The Poincaré-Birkhoff theorem is a fundamental result in the areas of fixed point theory and dynamical systems, as well as in their applications to differential equations. General presentations can be found in [28, 35, 37]. There is a large literature on the subject and certain subtle and delicate points related to some controversial extensions of the theorem have been settled only in recent years (see [29, 33, 46]). In the applications to the study of periodic non-autonomous planar Hamiltonian systems, the map Ψ is often the Poincaré map (or one of its iterates). In this situation the condition of boundary invariance is usually not satisfied, or very difficult to prove: as a consequence, variants of the Poincaré-Birkhoff theorem in which the hypothesis $(PB1)$ is not required turn out to be quite useful for the applications (see [14] for a general discussion on this topic). As a step in this direction we present the next result, following from Ding in [18].

Theorem 2. *Let $\Psi : D[R] \to \Psi(D[R]) \subseteq \mathbb{R}^2$ be an area preserving homeomorphism with $\Psi(0) = 0$ and such that the twist condition $(PB2)$ holds. Then Ψ has at least two fixed points z_1, z_2 in the interior of A and $\mathscr{J}(\theta, \rho) = j$ for $\Pi(\theta, \rho) = z_i$.*

The proof in [18] (see also [17, Appendix]) relies on the Jacobowitz version of the Poincaré-Birkhoff theorem for a pointed topological disk [25, 26] which was corrected in [29], since the result is true for strictly star-shaped pointed disks and not valid in general, as shown by a counterexample in the same article. Another (independent) proof of Theorem 2 was obtained by Rebelo in [46], who brought the proof back to that of Theorem 1 and thus to the "safe" version of Brown and Neumann [7]. Other versions of the Poincaré-Birkhoff theorem giving Theorem 2 as a corollary can be found in [23, 24, 32, 45] (see also [20, Introduction] for a general discussion about these delicate aspects). For Poincaré maps associated with Hamiltonian systems there is a much more general version of the theorem due to Fonda and Ureña in [21, 22], which holds in higher dimension, too.

In [15, 16], Ding proposed a variant of the Poincaré-Birkhoff theorem, by introducing the concept of "bend-twist map". Given a continuous map $\Psi : A \to \Psi(A) \subseteq \mathbb{R}^2 \setminus \{0\}$, which admits a lifting $\tilde{\Psi}$ as in (1), we define

$$\Upsilon(\theta, \rho) := \mathscr{R}(\theta, \rho) - \rho.$$

We call Ψ a *bend-twist map* if it Ψ satisfies the twist condition and Υ changes its sign on a non-contractible Jordan closed curve Γ contained in the set of points in the interior of A where $\mathscr{J} = j$. The original treatment was given in [15] for analytic maps. There are extensions to continuous maps as well [43, 44]. The bend-twist map condition is difficult to check in practice, due to the lack of information about the curve Γ (which, in the non-analytic case, may not even be a curve). For this reason, one can rely on the following corollary [15, Corollary 7.3] which also follows from the Poincaré-Miranda theorem (as observed in [43]).

Theorem 3. *Let $\Psi : A = A[r, R] \to \Psi(A) \subseteq \mathbb{R}^2 \setminus \{0\}$ be a continuous map such that the twist condition $(PB2)$ holds. Suppose that there are two disjoint arcs α, β contained in A, connecting the inner with the outer boundary of the annulus and such that*

$$\Upsilon > 0 \text{ on } \alpha \text{ and } \Upsilon < 0 \text{ on } \beta. \tag{$BT1$}$$

Then Ψ has at least two fixed points z_1, z_2 in the interior of A and $\mathscr{J}(\theta, \rho) = j$ for $\Pi(\theta, \rho) = z_i$.

A simple variant of the above theorem considers $2n$ pairwise disjoint simple arcs α_i and β_i (for $i = 1, \dots, n$) contained in A and connecting the inner with the outer boundary. We label these arcs in cyclic order so that each β_i is between α_i and α_{i+1} and each α_i is between β_{i-1} and β_i (with $\alpha_{n+1} = \alpha_1$ and $\beta_0 = \beta_n$) and suppose that

$$\Upsilon > 0 \text{ on } \alpha_i \text{ and } \Upsilon < 0 \text{ on } \beta_i\,, \quad \forall\, i = 1, \dots, n. \tag{BTn}$$

Then Ψ has at least $2n$ fixed points z_i in the interior of A and $\mathscr{J}(\theta, \rho) = j$ for $\Pi(\theta, \rho) = z_i$. These results also apply in the case of a topological annulus (namely, a compact planar set homeomorphic to A) and do not require that Ψ is area-preserving and also the assumption of Ψ being a homeomorphism is not required, as continuity is enough. Moreover, since the fixed points are obtained in regions with index ± 1, the results are robust with respect to small (continuous) perturbations of the map Ψ.

A special case in which condition $(BT1)$ holds is when $\Psi(\alpha) \in D(r)$ and $\Psi(\beta) \in \mathbb{R}^2 \setminus D[R]$, namely, the annulus A, under the action of the map Ψ, is not only twisted, but also strongly stretched, in the sense that there is a portion of the annulus around the curve α which is pulled inward near the origin inside the disc $D(r)$, while there is a portion of the annulus around the curve β which is pushed outside the disc $D[R]$. This special situation where a strong bend and twist occur is reminiscent of the geometry of the *Smale horseshoe maps* [36, 48] and, indeed, we will show how to enter in a variant of the theory of *topological horseshoes* in the sense of Kennedy and Yorke [27]. To this aim, we recall a few definitions which are useful for the present setting. By a *topological rectangle* we mean a subset $\mathscr{R}$ of the plane which is homeomorphic to the unit square. Given an arbitrary topological rectangle $\mathscr{R}$ we can define an orientation, by selecting two disjoint compact arcs on its boundary. The union of these arcs is denoted by $\mathscr{R}^-$ and the pair $\widehat{\mathscr{R}} := (\mathscr{R}, \mathscr{R}^-)$ is called an *oriented rectangle*. Usually the two components of $\mathscr{R}^-$ are labelled as the left and the right sides of $\widehat{\mathscr{R}}$. Given two oriented rectangles $\widehat{\mathscr{A}}$, $\widehat{\mathscr{B}}$, a continuous map Ψ and a compact set $H \subseteq \mathrm{dom}(\Psi) \cap \mathscr{A}$, the notation $(H, \Psi) : \widehat{\mathscr{A}} \Bumpeq\!\!\!\!\!\!\longrightarrow \widehat{\mathscr{B}}$ means that the following "stretching along the paths" (SAP) property is satisfied: *any path γ, contained in $\mathscr{A}$ and joining the opposite sides of $\mathscr{A}^-$, contains a sub-path σ in H such that the image of σ through Ψ is a path contained in $\mathscr{B}$ which connects the opposite sides of $\mathscr{B}^-$*. We also write $\Psi : \widehat{\mathscr{A}} \Bumpeq\!\!\!\!\!\!\longrightarrow \widehat{\mathscr{B}}$ when $H = \mathscr{A}$. By a path γ we mean a continuous map defined on a compact interval. When, loosely speaking, we say that a path is contained in a given set we actually refer to its image $\bar{\gamma}$. Sometimes it will be useful to consider a relation of the form $\Psi : \widehat{\mathscr{A}} \Bumpeq\!\!\!\!\!\!\longrightarrow^{k} \widehat{\mathscr{B}}$, for $k \geq 2$ a positive integer, which means that there are at least k compact subsets $H_1, \ldots, H_k$ of $\mathscr{A}$ such that $(H_i, \Psi) : \widehat{\mathscr{A}} \Bumpeq\!\!\!\!\!\!\longrightarrow \widehat{\mathscr{B}}$ for all $i = 1, \ldots, k$. From the results in [40, 41] we have that Ψ has a fixed point in H whenever $(H, \Psi) : \widehat{\mathscr{R}} \Bumpeq\!\!\!\!\!\!\longrightarrow \widehat{\mathscr{R}}$. If for a rectangle $\mathscr{R}$ we have that $\Psi : \widehat{\mathscr{R}} \Bumpeq\!\!\!\!\!\!\longrightarrow^{k} \widehat{\mathscr{R}}$, for $k \geq 2$, then Ψ has at least k fixed points in $\mathscr{R}$. In this latter situation, one can also prove the presence of chaotic-like dynamics of coin-tossing type (this will be briefly discussed later).

The aim of this paper is to analyze, under these premises, the second order scalar equation of Duffing type

$$x'' + q(t)g(x) = 0 \qquad (DE)$$

with $q(t)$ being a periodic sign-changing weight. The prototypical nonlinearity we consider is a function which changes sign at zero and is bounded only on one-side, such as $g(x) = -1 + \exp(x)$. We prove the presence of periodic solutions coming in pairs (Theorem 4 in Sect. 2, following the Poincaré-Birkhoff theorem) or coming in quadruplets (Theorem 5 in Sect. 2, following bend-twist maps and SAP techniques), the latter depending on the intensity of the negative part of $q(\cdot)$.

2 Statement of the Main Results

We express (DE) as a sign-indefinite nonlinear first order planar systems of the form

$$x' = y, \qquad y' = -a_{\lambda,\mu}(t)g(x). \tag{2}$$

Throughout the article, we suppose that $g : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz continuous function satisfying the following assumptions:

$$g(0) = 0, \quad g(x)x > 0 \text{ for all } x \neq 0, \quad g_0 := \liminf_{|x|\to 0} \frac{g(x)}{x} > 0. \tag{C_0}$$

We also suppose that *at least one* of the two following conditions holds:

$$(g_-) \quad g \text{ is bounded on } \mathbb{R}^-, \qquad (g_+) \quad g \text{ is bounded on } \mathbb{R}^+.$$

The weight function $q(t) := a_{\lambda,\mu}(t)$ is defined starting from a T-periodic sign-changing map $a : \mathbb{R} \to \mathbb{R}$ by setting

$$a_{\lambda,\mu}(t) = \lambda a^+(t) - \mu a^-(t), \qquad \lambda, \mu > 0,$$

where $a^+ := (a + |a|)/2$ is the positive part of $a(\cdot)$ and $a^- := a^+ - a$ is the negative one. Given an interval I, we denote by $a \succ 0$ on I the condition $a(t) \geq 0$ for almost every $t \in I$ with $a > 0$ on a subset of I of positive measure. Similarly, $a \prec 0$ on I means that $-a \succ 0$ on I. We suppose that, in a period, the weight function $a(t)$ displays one positive hump followed by one negative hump, i.e. there are t_0 and $T_1 \in]0, T[$ such that

$$a \succ 0 \quad \text{on } [t_0, t_0 + T_1] \quad \text{and} \quad a \prec 0 \quad \text{on } [t_0 + T_1, t_0 + T].$$

Due to the T-periodicity of the weight function, it is not restrictive to take $t_0 = 0$ and we shall assume it for the rest of the paper. As for the regularity of the weight function, we suppose that $a(\cdot)$ is continuous or piecewise-continuous (more general Carathéodory assumptions could be considered, too).

We consider the Poincaré map associated with system (2), namely

$$\Phi_{t_0}^{t}(z) := (x(t; t_0, z), y(t; t_0, z))$$

where $(x(\cdot\,; t_0, z), y(\cdot\,; t_0, z))$ is the solution of (2) satisfying the initial condition $z = (x(t_0), y(t_0))$ and set $\Phi(z) := \Phi_0^T(z)$. Since (2) has a Hamiltonian structure, the associated Poincaré map is an area-preserving homeomorphism, defined on a open set $\Omega := \mathrm{dom}\Phi \subseteq \mathbb{R}^2$, with $(0, 0) \in \Omega$. In view of the Introduction, a possible method to prove the existence (and multiplicity) of T-periodic solutions makes use of the Poincaré-Birkhoff theorem. Accordingly, we look for a suitable annulus around the origin with radii $0 < r_0 < R_0$ such that for some $\mathfrak{a} < \mathfrak{b}$ the twist condition

$$\mathrm{rot}_z(T) > \mathfrak{b} \ \forall z : ||z|| = r_0, \quad \mathrm{rot}_z(T) < \mathfrak{a} \ \forall z : ||z|| = R_0 \qquad (TC)$$

holds, where $\mathrm{rot}_z(T)$ is the rotation number on the interval $[0, T]$ associated with $z \in \mathbb{R}^2 \setminus \{(0, 0)\}$. In this setting, a standard definition of the rotation number is given by $\mathrm{rot}_z(T) := \mathrm{rot}_z(0, T)$, where

$$\mathrm{rot}_z(t_1, t_2) := \frac{1}{2\pi} \int_{t_1}^{t_2} \frac{y(t)^2 + a_{\lambda,\mu}(t)x(t)g(x(t))}{x^2(t) + y^2(t)} dt, \qquad (3)$$

being $(x(t), y(t))$ the solution of (2) with $(x(t_1), y(t_1)) = z \neq (0, 0)$. Notice that in (3) the angular displacement is positive when the rotations around the origin are performed in the clockwise sense.

Under these assumptions, the Poincaré-Birkhoff theorem, in the version of [46, Corollary 2], guarantees that for each integer $j \in [\mathfrak{a}, \mathfrak{b}]$, there exist *at least two* T-periodic solutions of system (2), having j as associated rotation number. It turns out that these solutions have precisely $2j$ simple transversal crossings with the y-axis in the interval $[0, T[$ (see, for instance, [30, Theorem A]). Equivalently, for any periodic solution $(x(t), y(t))$, we have that x has precisely $2j$ simple zeros in the interval $[0, T[$.

We stress that, to apply this approach, the Poincaré map must be well defined on the annulus—actually, on the whole closed disc $D[R_0]$, that is $D[R_0] \subseteq \Omega$. As shown in [13], for the superlinear equation $x'' + q(t)x^{2n+1} = 0$ (with $n \geq 1$), even for a positive weight $q(t)$ the global existence of the trajectories is not guaranteed, due to the presence of solutions which blow-up in finite time with infinitely many winds around the origin. In our case, the boundedness assumption at infinity, given by one among (g_-) or (g_+), prevents such highly oscillatory phenomenon and guarantees the continuability on $[0, T_1]$. In the time intervals where the weight function is negative, we cannot prevent blow-up phenomena (see [8]) unless we impose some growth restrictions on the vector field (for instance, assuming both (g_-) and (g_+)).

At this point, if we are willing to assume the global continuability for the solutions of (2), the following result can be stated.

Theorem 4. *Assume* (C_0) *and* (g_-) *or* (g_+)*. Then, for each positive integer* k*, there exists* $\Lambda_k > 0$ *such that for each* $\lambda > \Lambda_k$ *and* $j = 1, \dots, k$*, the equation* (DE) *has at least* two *T-periodic solutions having exactly* $2j$*-zeros in the interval* $[0, T[$.

Notice that no condition on the parameter $\mu > 0$ is required. On the other hand, we are forced to suppose the global continuability of the solutions. The next result overcomes the difficulties related to the Poincaré-Birkhoff approach, by using a different fixed point theorem which requires μ to be sufficiently large.

Theorem 5. *Assume* (C_0) *and* (g_-) *or* (g_+)*. Then, for each positive integer* k*, there exists* $\Lambda_k > 0$ *such that for each* $\lambda > \Lambda_k$ *there exists* $\mu^* = \mu^*(\lambda)$ *such that for each* $\mu > \mu^*$ *and* $j = 1, \dots, k$*, the equation* (DE) *has at least* four *T-periodic solutions having exactly* $2j$*-zeros in the interval* $[0, T[$.

In [19] the general proofs of Theorem 4 and Theorem 5 are given directly for a class of planar systems including (2). Theorem 4 is related to a previous work by Boscaggin [5], dealing with subharmonic solutions. Concerning Theorem 5, we propose in the next section a different proof in the special case of a stepwise weight function. The simplified form of the weight allows us to display the geometric features of the problem and to provide more detailed information on the distribution of the zeros.

3 Proofs. A Simplified Geometric Framework

We focus on the particular case in which $g : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz continuous function satisfying (C_0) along with (g_-). A possible choice could be $g(x) = e^x - 1$, but we stress that we do not ask for g to be unbounded on $\mathbb{R}^+$. Recalling the choice of $q(t)$, we rewrite (DE) as

$$x'' + a_{\lambda,\mu}(t)g(x) = 0. \tag{4}$$

In order to illustrate quantitatively the main ideas of the proof we choose a stepwise T-periodic function $a(\cdot)$ which takes value $a(t) = 1$ on an interval of length T_1 and value $a(t) = -1$ on a subsequent interval of length $T_2 = T - T_1$, so that $a_{\lambda,\mu}$ is defined as

$$a_{\lambda,\mu}(t) = \begin{cases} \lambda & \text{for } t \in [0, T_1[\\ -\mu & \text{for } t \in [T_1, T_1 + T_2[\end{cases} \qquad T_1 + T_2 = T. \tag{5}$$

With this particular choice of $a(t)$, the planar system associated with (4) turns out to be a periodic switched system [2]. Such kind of systems are widely studied in control theory.

For our analysis we first take into account the interval of positivity for the weight, where (2) becomes

$$x' = y, \qquad y' = -\lambda g(x). \tag{6}$$

For this system the origin is a local center, which is global if $\mathscr{G}(x) \to +\infty$ as $x \to \pm\infty$, where $\mathscr{G}(x)$ is the primitive of $g(x)$ such that $\mathscr{G}(0) = 0$. The associated energy function is given by

$$E_1(x, y) := \frac{1}{2}y^2 + \lambda\mathscr{G}(x).$$

For any constant c with $0 < c < \min\{\mathscr{G}(-\infty), \mathscr{G}(+\infty)\}$, the level line of (6) of positive energy λc is a closed orbit Γ which intersects the x-axis in the phase-plane at two points $(x_-, 0)$ and $(x_+, 0)$ such that $x_- < 0 < x_+$, and $c := \mathscr{G}(x_-) = \mathscr{G}(x_+) > 0$. We call $\tau(c)$ the period of Γ, which is given by

$$\tau(c) = \tau^+(c) + \tau^-(c),$$

where

$$\tau^+(c) := \sqrt{\frac{2}{\lambda}} \int_0^{x_+} \frac{d\xi}{\sqrt{(c - \mathscr{G}(\xi))}}, \quad \tau^-(c) := \sqrt{\frac{2}{\lambda}} \int_{x_-}^{0} \frac{d\xi}{\sqrt{(c - \mathscr{G}(\xi))}}$$

The maps $c \mapsto \tau^{\pm}(c)$ are continuous. To proceed with our discussion, we suppose that $\mathscr{G}(-\infty) \leq \mathscr{G}(+\infty)$ (the other situation can be treated symmetrically). Then $\tau^-(c) \to +\infty$ as $c \to \mathscr{G}(-\infty)$ (this follows from the fact that $g(x)/x$ goes to zero as $x \to -\infty$, see [38]). We can couple this result with an estimate near the origin

$$\limsup_{c \to 0^+} \tau(c) \leq 2\pi / \sqrt{\lambda g_0}$$

which follows from classical and elementary arguments.

Proposition 1. *For each $\lambda > 0$, the time-mapping τ associated with system (6) is continuous and its range includes the interval $]2\pi/\sqrt{\lambda g_0}, +\infty[$.*

Showing the monotonicity of the whole time-map $\tau(c)$ is, in general, a difficult task. However, for the exponential case $g(x) = e^x - 1$ this has been proved in [11] (see also [10]).

On the interval of negativity of $a_{\lambda,\mu}(t)$, system (2) becomes

$$x' = y, \qquad y' = \mu g(x), \tag{7}$$

with $g(x)$ as above. For this system the origin is a global saddle with unbounded stable and unstable manifolds contained in the zero level set of the energy

$$E_2(x, y) := \frac{1}{2} y^2 - \mu \mathscr{G}(x).$$

If we start from a point $(0, y_0)$ with $y_0 > 0$ we can explicitly evaluate the blow-up time as follows. First of all we compute the time needed to reach the level $x = \kappa > 0$ along the trajectory of (7), which is the curve of fixed energy $E_2(x, y) = E_2(0, y_0)$ with $y > 0$. Equivalently, we have

$$y = x' = \sqrt{y_0^2 + 2\mu \mathscr{G}(x)}$$

from which

$$t = \int_0^{\kappa} \frac{dx}{\sqrt{y_0^2 + 2\mu \mathscr{G}(x)}}$$

follows. Therefore, the blow-up time is given by

$$T(y_0) = \int_0^{+\infty} \frac{dx}{\sqrt{y_0^2 + 2\mu \mathscr{G}(x)}}.$$

Standard theory guarantees that if the Keller-Osserman condition

$$\int^{+\infty} \frac{dx}{\sqrt{\mathscr{G}(x)}} < +\infty \tag{8}$$

holds, then the blow-up time is always finite and $T(y_0) \searrow 0$ for $y_0 \nearrow +\infty$. On the other hand, $T(y_0) \nearrow +\infty$ for $y_0 \searrow 0^+$. Hence there exists $\bar{y} > 0$ such that $T(y_0) > T_2$ for $y_0 \in]0, \bar{y}[$ and hence there is no blow-up in $[T_1, T]$.

If we start with null derivative, i.e. from a point $(x_0, 0)$, then similar calculations return

$$t = \int_{x_0}^{\kappa} \frac{dx}{\sqrt{2\mu(\mathscr{G}(x) - \mathscr{G}(x_0))}}$$

and, since $\mathscr{G}(x) - \mathscr{G}(x_0) \sim g(x_0)(x - x_0)$ for $|x - x_0| \ll 1$, the improper integral at x_0 is finite. Therefore, the blow-up time is given by

$$T(x_0) = \int_{x_0}^{+\infty} \frac{dx}{\sqrt{2\mu(\mathscr{G}(x) - \mathscr{G}(x_0))}}.$$

If (8) is satisfied, then the blow-up time is always finite. Moreover, $T(x_0) \to +\infty$ as $x_0 \to 0^+$. A similar but more refined result can be found in [39, Lemma 3].

Now we describe how to obtain Theorem 4 and Theorem 5 for system (2) in the special case of a T-periodic stepwise function as in (5). As we already observed, due to the special form of the weight function, equation (2) is a periodic switched system and therefore its associated Poincaré map Φ on the interval $[0, T]$ splits as $\Phi = \Phi_2 \circ \Phi_1$ where Φ_1 is the Poincaré map on the interval $[0, T_1]$ associated with system (6) and Φ_2 is the Poincaré map on the interval $[0, T_2]$ associated with system (7).

(I). Proof of Theorem 4 for the Stepwise Weight

Proof. We start by selecting a closed orbit Γ^0 near the origin of (6) at a level energy λc_0 and fix λ sufficiently large, say $\lambda > \Lambda_k$, so that in view of Proposition 1

$$\tau(c_0) < \frac{T_1}{k+1}. \tag{9}$$

Next, for the given (fixed) λ, we consider a second energy level λc_1 with $c_1 > c_0$ such that

$$\tau^{-}(c_1) > 2T_2 \tag{10}$$

and denote by Γ^1 the corresponding closed orbit. Let also

$$\mathscr{A} := \{(x, y) : 2\lambda c_0 \leq y^2 + 2\lambda\mathscr{G}(x) \leq 2\lambda c_1\}$$

be the planar annular region enclosed between Γ^0 and Γ^1. If we assume that the Poincaré map Φ_2 is defined on $\mathscr{A}$, then the complete Poincaré map Φ associated with system (2) is a well defined area-preserving homeomorphism of the annulus $\mathscr{A}$ onto its image $\Phi(\mathscr{A}) = \Phi_2(\mathscr{A})$. In fact the annulus is invariant under the action of Φ_1.

During the time interval $[0, T_1]$, each point $z \in \Gamma^0$ performs $\lfloor T_1/\tau(c_0) \rfloor$ complete turns around the origin in the clockwise sense. This implies that

$$\mathrm{rot}_z(0, T_1) \geq \left\lfloor \frac{T_1}{\tau(c_0)} \right\rfloor, \quad \forall z \in \Gamma^0.$$

On the other hand, from [6, Lemma 3.1] we know that

$$\mathrm{rot}_z(T_1, T) = \mathrm{rot}_z(0, T_2) > -\frac{1}{2}, \quad \forall z \neq (0, 0).$$

We conclude that $\mathrm{rot}_z(T) > k$, for all $z \in \Gamma^0$.

During the time interval $[0, T_1]$, each point $z \in \Gamma^1$ is unable to complete a full revolution around the origin, because the time needed to cross either the second or the third quadrant is larger than T_1. Using this information in connection to the fact that the first and the third quadrants are positively invariant for the flow associated with (7), we find that $\mathrm{rot}_z(T) < 1$, for all $z \in \Gamma^1$.

Thus we have condition (TC) matched with $\mathfrak{b} = k$ and $\mathfrak{a} = 1$. An application of the Poincaré-Birkhoff fixed point theorem [46] (this time for a topological annulus with strictly star-shaped boundaries) guarantees for each $j = 1, \ldots, k$ the existence of at least two fixed points $u_j = (u_x^j, u_y^j)$, $v_j = (v_x^j, v_y^j)$ of the Poincaré map, with u_j, v_j in the interior of $\mathscr{A}$ and such that $\mathrm{rot}_{u_j}(T) = \mathrm{rot}_{v_j}(T) = j$. This in turns implies the existence of at least two T-periodic solutions of Eq. (4) with $x(\cdot)$ having exactly $2j$-zeros in the interval $[0, T]$. □

In this manner, we have proved Theorem 4 for system (2) in the special case of a stepwise weight function $a_{\lambda,\mu}$ as in (5). Notice that no assumption on $\mu > 0$ is required. On the other hand, we have to suppose that Φ_2 is globally defined on $\mathscr{A}$.

Remark 1. From (9) and the formulas for the period τ it is clear that assuming T_1 fixed and λ large is equivalent to suppose λ fixed and T_1 large. This also follows from general considerations concerning the fact that equation $x'' + \lambda g(x) = 0$ is equivalent to $u'' + \varepsilon^2 \lambda g(u) = 0$ for $u(\xi) := x(\varepsilon\xi)$. ◁

(II). An Intermediate Step

Now we show how to improve the previous result if we add the condition that μ is sufficiently large. First of all, we take Γ^0 and Γ^1 as before and $\lambda > \Lambda_k$ in order to produce the desired twist for Φ at the boundary of $\mathscr{A}$. Then we observe that the derivative of the energy E_1 along the trajectories of system (7) is given by $(\lambda + \mu) y g(x)$, so it increases on the first and the third quadrant and decreases on the second and the fourth. Hence, if μ is sufficiently large, we can find four arcs $\varphi_i \subseteq \mathscr{A}$, each one in the open i-th quadrant, with φ_i joining Γ^0 and Γ^1 such that $\Phi_2(\varphi_i)$ is outside the region bounded by Γ^1 for $i = 1, 3$ and $\Phi_2(\varphi_i)$ is inside the region bounded by Γ^0 for $i = 2, 4$. The corresponding position of $\mathscr{A}$ and $\Phi_2(\mathscr{A})$ is illustrated in Fig. 1.

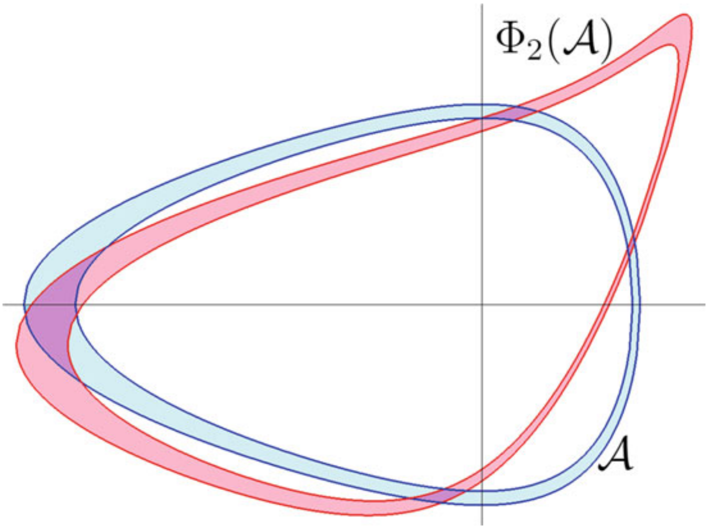


Fig. 1 A possible configuration of $\mathscr{A}$ and $\Phi_2(\mathscr{A})$. The example is obtained for $g(x) = -1 + \exp x$, $\lambda = \mu = 0.1$ and $T_2 = 1$. The inner and outer boundary Γ^0 and Γ^1 of the annulus $\mathscr{A}$ are the energy level lines $E_1(x, y) = E_1(2, 0)$ and $E_1(x, y) = E_1(2.1, 0)$. To produce this geometry, the value of T_1 is not relevant because the annulus is invariant for system (6). Since $\tau(c_0) < \tau(c_1)$, to have a desired twist condition, we need to assume T_1 large enough

At this point, we enter in the setting of bend-twist maps. The arcs $\Phi_1^{-1}(\varphi_i)$ divide $\mathscr{A}$ into four regions, homeomorphic to rectangles. The boundary of each of these regions can be split into two opposite sides contained in Γ^0 and Γ^1 and two other opposite sides made by $\Phi_1^{-1}(\varphi_i)$ and $\Phi_1^{-1}(\varphi_{i+1})$ (in cyclic order). On Γ^0 and Γ^1 we have the previously proved twist condition on the rotation numbers, while on the other two sides we have $E_1(\Phi(P)) > E_1(P)$ for $P \in \Phi_1^{-1}(\varphi_i)$ with $i = 1, 3$ and $E_1(\Phi(P)) < E_1(P)$ for $P \in \Phi_1^{-1}(\varphi_i)$ with $i = 2, 4$. Thus, using the Poincaré-Miranda theorem, we obtain the existence of at least one fixed point of the Poincaré map Φ in the interior of each of these regions. In this manner, under an additional hypothesis of the form $\mu > \mu^*(\lambda)$, we improve Theorem 4 (for system (2) and again in the special case of a stepwise weight), finding at least four solutions with a given rotation number j for $j = 1, \dots, k$. On the other hand, we still suppose that Φ_2 is globally defined on $\mathscr{A}$. The version of the bend-twist map theorem that we apply here is robust for small perturbations of the Poincaré map, therefore the result holds also for some non-Hamiltonian systems whose vector field is close to that of (4). □

(III). Proof of Theorem 5 for the Stepwise Weight

Proof. First of all, we start with the same construction as in **(I)** and choose Γ^0, $\lambda > \Lambda_k$ according to (9) and Γ^1 so that (10) is satisfied. Consistently with the previously introduced notation, we take

$$x_-^1 < x_-^0 < 0 < x_+^0 < x_+^1, \quad \text{with } \mathscr{G}(x_-^i) = c_i = \mathscr{G}(x_+^i), \ i = 0, 1.$$

Notice that the closed curves Γ^i intersect the coordinate axes at the points $(x^i_\pm, 0)$ and $(0, \pm\sqrt{2\lambda c_i})$. Next we choose $x^\mu_\pm$ and y_0 with

$$x^0_- < x^\mu_- < 0 < x^\mu_+ < x^0_+, \quad \text{and } 0 < y_0 < \sqrt{2\lambda c_0}$$

and define the orbits

$$\mathscr{X}_\pm := \gamma(x^\mu_\pm, 0), \quad \mathscr{Y}_\pm := \gamma(0, \pm y_0),$$

where we denote by $\gamma(P)$ the complete orbit of the system passing through the point $P \in \mathbb{R}^2$.

Setting

$$\mathscr{T}(\mathscr{X}_\pm) := \pm 2 \int_{x^\mu_\pm}^{x^1_\pm} \frac{dx}{\sqrt{2\mu(\mathscr{G}(x) - \mathscr{G}(x^\mu_\pm))}}, \quad \mathscr{T}(\mathscr{Y}) := \int_{x^1_-}^{x^1_+} \frac{dx}{\sqrt{y_0^2 + 2\mu\mathscr{G}(x)}}$$

we tune the values $x^\mu_\pm$, y_0 and μ so that

$$\max\{\mathscr{T}(\mathscr{X}_\pm), \mathscr{T}(\mathscr{Y})\} < T_2.$$

Clearly, given the other parameters, we can always choose μ sufficiently large, say $\mu > \mu^*$, so that the above condition is satisfied.

Finally, we introduce the stable and unstable manifolds, W^s and W^u, for the origin as saddle point of system (7). More precisely, we define the sets

$$W^s_+ := \{(x, y) : E_2(x, y) = 0, x > 0, y < 0\}, \ W^s_- := \{(x, y) : E_2(x, y) = 0, x < 0, y > 0\},$$

$$W^u_+ := \{(x, y) : E_2(x, y) = 0, x > 0, y > 0\}, \ W^u_- := \{(x, y) : E_2(x, y) = 0, x < 0, y < 0\},$$

so that $W^s = W^s_- \cup W^s_+$ and $W^u = W^u_- \cup W^u_+$. The resulting configuration is illustrated in Fig. 2.

The closed trajectories Γ^0, Γ^1 together with $\mathscr{X}_\pm$, $\mathscr{Y}_\pm$, $W^s_\pm$ and $W^u_\pm$ determine eight regions that we denote by $\mathscr{A}_i$ and $\mathscr{B}_i$ for $i = 1, \dots, 4$, as in Fig. 3.

Each of the regions $\mathscr{A}_i$ and $\mathscr{B}_i$ is homeomorphic to the unit square and thus is a topological rectangle. In this setting, we give an orientation to $\mathscr{A}_i$ by choosing $\mathscr{A}_i^- := \mathscr{A}_i \cap (\Gamma^0 \cup \Gamma^1)$. We take as $\mathscr{B}_i^-$ the closure of $\partial\mathscr{B}_i \setminus (\Gamma^0 \cup \Gamma^1)$.

We can now apply a result in the framework of the theory of topological horseshoes as presented in [42] and [31]. Indeed, by the previous choice of $\lambda > \Lambda_k$ we obtain that

$$\Phi_1 : \widehat{\mathscr{A}_i} \Bumpeq\!\!\!\!\longrightarrow^k \widehat{\mathscr{B}_i}, \quad \forall i = 1, \dots, 4,$$

On the other hand, from $\mu > \mu^*$ it follows that

$$\Phi_2 : \widehat{\mathscr{B}_i} \Bumpeq\!\!\!\!\longrightarrow \widehat{\mathscr{A}_i}, \quad \forall i = 1, \dots, 4.$$

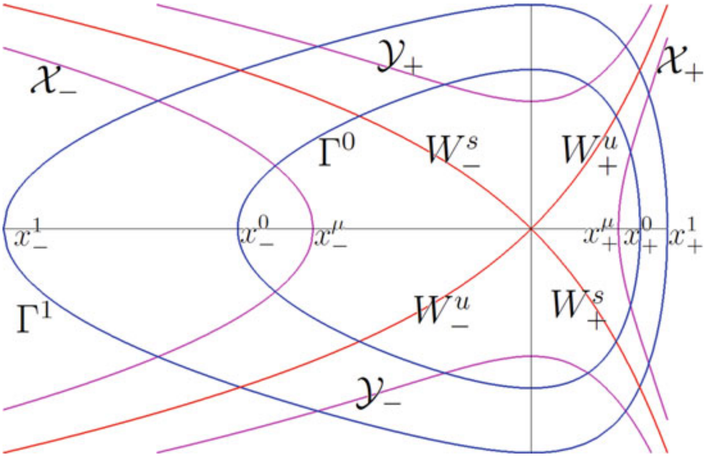


Fig. 2 The present figure shows the appropriate overlapping of the phase-portraits of systems (6) and (7)

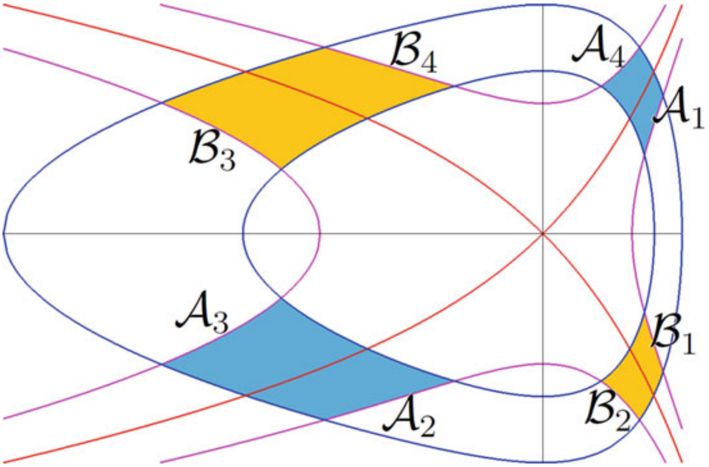


Fig. 3 The present figure shows the regions $\mathscr{A}_i$ and $\mathscr{B}_i$. We have labelled the regions following a clockwise order, which is useful from the point of view of the dynamics

Then [42, Theorem 3.1] (see also [31, Theorem 2.1]) ensures the existence of at least k fixed points for $\Phi = \Phi_2 \circ \Phi_1$ in each of the regions $\mathscr{A}_i$. This, in turns, implies the existence of $4k$ T-periodic solutions for system (2).

Such solutions are topologically different and can be classified, as follows: for each $j = 1, \dots, k$ there is a solution (x, y) with

- ∘ $(x(0), y(0)) \in \mathscr{A}_1$ with $x(t)$ having $2j$ zeros in $]0, T_1[$ and strictly positive in $[T_1, T]$;
- ∘ $(x(0), y(0)) \in \mathscr{A}_2$ with $x(t)$ having $2j - 1$ zeros in $]0, T_1[$ and one zero in $]T_1, T[$;

- ∘ $(x(0), y(0)) \in \mathscr{A}_3$ with $x(t)$ having $2j$ zeros in $]0, T_1[$ and strictly negative in $[T_1, T]$;
- ∘ $(x(0), y(0)) \in \mathscr{A}_4$ with $x(t)$ having $2j - 1$ zeros in $]0, T_1[$ and one zero in $]T_1, T[$.

In conclusion, for each $j = 1, \dots, k$ we find at least four T-periodic solutions having precisely $2j$-zeros in $[0, T[$. □

Remark 2. Having assumed that g is bounded on $\mathbb{R}^-$, we can also prove the existence of a T-periodic solution with $(x(0), y(0)) \in \mathscr{A}_3$ and such that $x(t) < 0$ for all $t \in [0, T]$ while $y(t) = x'(t)$ has two zeros in $[0, T[$. Moreover, the results from [31, 42] guarantee also that each of the regions $\mathscr{A}_i$ contains a compact invariant set where Φ is chaotic in the sense of Block and Coppel (see [1, 34]). At last, we also mention that the result (from Theorem 5) is robust with respect to small perturbations. In particular, it applies to a perturbed Hamiltonian system of the form

$$x' = y + F_1(t, x, y, \varepsilon), \quad y' = -a_{\lambda,\mu} g(x) + F_2(t, x, y, \varepsilon) \tag{11}$$

with $F_1, F_2 \to 0$ as $\varepsilon \to 0$, uniformly in t, and for (x, y) on compact sets. Observe that system (11) has not necessarily a Hamiltonian structure and therefore it is no more guaranteed that the associated Poincaré map is area-preserving. ◁

Remark 3. We further observe that, for Eq. (4) the same results hold if condition (g_-) is relaxed to

$$\lim_{x \to -\infty} \frac{g(x)}{x} = 0. \tag{12}$$

Under the same condition at infinity, four T-periodic solutions are obtained also in [6]. However, we stress that, the assumptions at the origin are completely different. Indeed, in [6] a one-sided superlinear condition in zero, of the form $g'(0^+) = 0$ or $g'(0^-) = 0$ was required. As a consequence, for λ large, one could prove the existence of four (or $4k$) T-periodic solutions with prescribed nodal properties which come in pair, namely two "small" and two "large". In our case, if in place of $g_0 > 0$ we assume $g'(0^+) = 0$ or $g'(0^-) = 0$, with the same approach we could prove the existence of eight (or $8k$) T-periodic solutions, four "small" and four "large". ◁

Remark 4. We conclude this note by observing that if we want to produce the same results for non-autonomous perturbations of the more general system

$$x' = h(y), \qquad y' = -g(x), \tag{13}$$

then we cannot replace (g_-) (or (g_+) with a weaker condition of the form of (12). In fact, a crucial step in our proof is to have a twist condition, that is a gap in the period between a fast orbit (like Γ^0) and slow one (like Γ^1). This is no more guaranteed for an autonomous system of the form (13) if $g(x)$ satisfies a sublinear condition at infinity as (12). Indeed, the slow decay of g at infinity could be compensated by a fast growth of h at infinity. In [12] the Authors provide examples of isochronous centers for planar Hamiltonian systems even in the case when one of the two components is sublinear at infinity. See [19] for perturbations of system (13) with a periodic sign-changing weight on the second equation. ◁

References

1. Aulbach, B., Kieninger, B.: On three definitions of chaos. Nonlinear Dyn. Syst. Theory **1**(1), 23–37 (2001)
2. Bacciotti, A.: Stability of switched systems: an introduction. In: Large-Scale Scientific Computing. Lecture Notes in Computer Science, vol. 8353, pp. 74–80. Springer, Heidelberg (2014)
3. Birkhoff, G.D.: Proof of Poincaré's geometric theorem. Trans. Am. Math. Soc. **14**(1), 14–22 (1913)
4. Bonino, M.: A topological version of the Poincaré-Birkhoff theorem with two fixed points. Math. Ann. **352**(4), 1013–1028 (2012)
5. Boscaggin, A.: Subharmonic solutions of planar Hamiltonian systems: a rotation number approach. Adv. Nonlinear Stud. **11**(1), 77–103 (2011)
6. Boscaggin, A., Zanolin, F.: Pairs of nodal solutions for a class of nonlinear problems with one-sided growth conditions. Adv. Nonlinear Stud. **13**(1), 13–53 (2013)
7. Brown, M., Neumann, W.D.: Proof of the Poincaré-Birkhoff fixed point theorem. Michigan Math. J. **24**(1), 21–31 (1977)
8. Burton, T., Grimmer, R.: On continuability of solutions of second order differential equations. Proc. Am. Math. Soc. **29**, 277–283 (1971)
9. Carter, P.H.: An improvement of the Poincaré-Birkhoff fixed point theorem. Trans. Am. Math. Soc. **269**(1), 285–299 (1982)
10. Chicone, C.: The monotonicity of the period function for planar Hamiltonian vector fields. J. Diff. Equat. **69**(3), 310–321 (1987)
11. Chow, S.-N., Wang, D.: On the monotonicity of the period function of some second order equations. Časopis Pěst. Mat. **111**(1), 14–25 (1986)
12. Cima, A., Gasull, A., Mañosas, F.: Period function for a class of Hamiltonian systems. J. Diff. Equat. **168**(1), 180–199 (2000). Special issue in celebration of Jack K. Hale's 70th birthday, Part 1 (Atlanta, GA/Lisbon, 1998)
13. Coffman, C.V., Ullrich, D.F.: On the continuation of solutions of a certain non-linear differential equation. Monatsh. Math. **71**, 385–392 (1967)
14. Dalbono, F., Rebelo, C.: Poincaré-Birkhoff fixed point theorem and periodic solutions of asymptotically linear planar Hamiltonian systems. Rend. Sem. Mat. Univ. Politec. Torino **60**(4), 233–263 (2003). 2002, Turin Fortnight Lectures on Nonlinear Analysis (2001)
15. Ding, T.: Approaches to the qualitative theory of ordinary differential equations. In: Dynamical Systems and Nonlinear Oscillations. Peking University Series in Mathematics, vol. 3. World Scientific Publishing Co. Pte. Ltd., Hackensack (2007)
16. Ding, T.-R.: The twist-bend theorem with an application. Adv. Math. (China) **41**(1), 31–44 (2012)
17. Ding, T.R., Zanolin, F.: Periodic solutions of Duffing's equations with superquadratic potential. J. Diff. Equat. **97**(2), 328–378 (1992)
18. Ding, W.Y.: Fixed points of twist mappings and periodic solutions of ordinary differential equations. Acta Math. Sinica **25**(2), 227–235 (1982)
19. Dondè, T., Zanolin, F.: Multiple periodic solutions for one-sided sublinear systems: a refinement of the Poincaré-Birkhoff approach. Topol. Methods Nonlinear Anal. **55**(2), 565–581 (2020)

20. Fonda, A., Sabatini, M., Zanolin, F.: Periodic solutions of perturbed Hamiltonian systems in the plane by the use of the Poincaré-Birkhoff theorem. Topol. Methods Nonlinear Anal. **40**(1), 29–52 (2012)
21. Fonda, A., Ureña, A.J.: A higher-dimensional Poincaré-Birkhoff theorem without monotone twist. C. R. Math. Acad. Sci. Paris **354**(5), 475–479 (2016)
22. Fonda, A., Ureña, A.J.: A higher dimensional Poincaré-Birkhoff theorem for Hamiltonian flows. Ann. Inst. H. Poincaré Anal. Non Linéaire **34**(3), 679–698 (2017)
23. Franks, J.: Generalizations of the Poincaré-Birkhoff theorem. Ann. of Math. 2 **128**(1), 139–151 (1988)
24. Franks, J.: Erratum to: "generalizations of the Poincaré-Birkhoff theorem". Ann. Math. (2) **128**(1), 139–151 (1988), mr0951509. Ann. Math. (2), **164**(3), 1097–1098 (2006)
25. Jacobowitz, H.: Periodic solutions of $x'' + f(x, t) = 0$ via the Poincaré-Birkhoff theorem. J. Diff. Equat. **20**(1), 37–52 (1976)
26. Jacobowitz, H.: Corrigendum: the existence of the second fixed point: a correction to "Periodic solutions of $x'' + f(x, t) = 0$ via the Poincaré-Birkhoff theorem". J. Diff. Equat. **20**(1), 37–52 (1976). J. Diff. Equat. **25**(1), S148–149 (1977)
27. Kennedy, J., Yorke, J.A.: Topological horseshoes. Trans. Am. Math. Soc. **353**(6), 2513–2530 (2001)
28. Le Calvez, P.: About Poincaré-Birkhoff theorem. Publ. Mat. Urug. **13**, 61–98 (2011)
29. Le Calvez, P., Wang, J.: Some remarks on the Poincaré-Birkhoff theorem. Proc. Am. Math. Soc. **138**(2), 703–715 (2010)
30. Margheri, A., Rebelo, C., Zanolin, F.: Maslov index, Poincaré-Birkhoff theorem and periodic solutions of asymptotically linear planar Hamiltonian systems. J. Diff. Equato **183**(2), 342–367 (2002)
31. Margheri, A., Rebelo, C., Zanolin, F.: Chaos in periodically perturbed planar Hamiltonian systems using linked twist maps. J. Diff. Equat. **249**(12), 3233–3257 (2010)
32. Marò, S.: Periodic solutions of a forced relativistic pendulum via twist dynamics. Topol. Methods Nonlinear Anal. **42**(1), 51–75 (2013)
33. Martins, R., Ureña, A.J.: The star-shaped condition on Ding's version of the Poincaré-Birkhoff theorem. Bull. Lond. Math. Soc. **39**(5), 803–810 (2007)
34. Medio, A., Pireddu, M.: Zanolin, F.: Chaotic dynamics for maps in one and two dimensions: a geometrical method and applications to economics. Int. J. Bifur. Chaos Appl. Sci. Eng. **19**(10), 3283–3309 (2009)
35. Meyer, K.R., Hall, G.R.: Introduction to Hamiltonian Dynamical Systems and the N-Body Poblem. Springer, New York (1992)
36. Moser, J.: Stable and Random Motions in Dynamical Systems. Princeton University Press, Princeton, University of Tokyo Press, Tokyo (1973). With special emphasis on celestial mechanics, Hermann Weyl Lectures, the Institute for Advanced Study, Princeton, N.J., Annals of Mathematics Studies, No. 77
37. Moser, J., Zehnder, E.J.: Notes on dynamical systems. Courant Lecture Notes in Mathematics, vol. 12. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence (2005)
38. Opial, Z.: Sur les périodes des solutions de l'équation différentielle $x'' + g(x) = 0$. Ann. Polon. Math. **10**, 49–72 (1961)
39. Papini, D., Zanolin, F.: A topological approach to superlinear indefinite boundary value problems. Topol. Methods Nonlinear Anal. **15**(2), 203–233 (2000)
40. Papini, D., Zanolin, F.: Fixed points, periodic points, and coin-tossing sequences for mappings defined on two-dimensional cells. Fixed Point Theory Appl. **2**, 113–134 (2004)
41. Papini, D., Zanolin, F.: On the periodic boundary value problem and chaotic-like dynamics for nonlinear Hill's equations. Adv. Nonlinear Stud. **4**(1), 71–91 (2004)
42. Pascoletti, A., Pireddu, M.,. Zanolin, F.: Multiple periodic solutions and complex dynamics for second order ODEs via linked twist maps. In: The 8th Colloquium on the Qualitative Theory of Differential Equations. Proceeding Colloquium on the Qualitative Theory of Differential Equations, vol. 8, pp. No. 14, 32. Electronic Journal of Qualitative Theory of Differential Equations, Szeged (2008)

43. Pascoletti, A., Zanolin, F.: A topological approach to bend-twist maps with applications. Int. J. Diff. Equat., Article no. 612041, 20 (2011)
44. Pascoletti, A., Zanolin, F.A.: Crossing lemma for annular regions and invariant sets with an application to planar dynamical systems. J. Math., Article no. 267393, 12 (2013)
45. Qian, D., Torres, P.J.: Periodic motions of linear impact oscillators via the successor map. SIAM J. Math. Anal. **36**(6), 1707–1725 (2005)
46. Rebelo, C.: A note on the Poincaré-Birkhoff fixed point theorem and periodic solutions of planar systems. Nonlinear Anal. **29**(3), 291–311 (1997)
47. Slaminka, E.E.: Removing index 0 fixed points for area preserving maps of two-manifolds. Trans. Am. Math. Soc. **340**(1), 429–445 (1993)
48. Smale, S.: Differentiable dynamical systems. Bull. Am. Math. Soc. **73**, 747–817 (1967)

Dynamical Models of Interrelation in a Class of Artificial Networks

Felix Sadyrbaev, Svetlana Atslega, and Eduard Brokan

Abstract The system of ordinary differential equations that models a type of artificial networks is considered. The system consists of a sigmoidal function that depends on linear combinations of the arguments minus the linear part. The linear combinations of the arguments are described by the regulatory matrix W. For the three-dimensional cases, several types of matrices W are considered and the behavior of solutions of the system is analyzed. The attractive sets are constructed for most cases. The illustrative examples are provided. The list of references consists of 12 items.

Keywords Gene regulatory networks · Dynamical systems · Artificial networks · Critical points · Attractors

Mathematics Subject Classification: 34C60 · 34D45 · 92B20

1 Introduction

This article is devoted to the study of attracting sets of some systems of ordinary differential equations that arise in the theory of artificial networks [11], genomic regulatory networks [5, 6] and appears as an auxiliary instrument used in the design of practically used networks [7]. We consider the system

F. Sadyrbaev (✉) · S. Atslega
Institute of Mathematics and Computer Science, University of Latvia, Rainis boul. 29, Riga, Latvia
e-mail: felix@latnet.lv

E. Brokan
Daugavpils University, Parades street 1, Daugavpils, Latvia

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_18

$$\begin{cases} \dfrac{dx_1}{dt} = \dfrac{1}{1+e^{-\mu_1(w_{11}x_1+w_{12}x_2+w_{13}x_n-\theta_1)}} - v_1x_1, \\ \dfrac{dx_2}{dt} = \dfrac{1}{1+e^{-\mu_2(w_{21}x_1+w_{22}x_2+w_{23}x_n-\theta_2)}} - v_2x_2, \\ \dfrac{dx_3}{dt} = \dfrac{1}{1+e^{-\mu_3(w_{31}x_1+w_{32}x_2+w_{33}x_3-\theta_3)}} - v_3x_3, \end{cases} \tag{1}$$

that can be considered as a model of interrelation in an artificial network, where each $x_i(t)$ is a characteristic of the dynamics of i-th node. Any equation consists of a nonlinear part minus the linear one. The linear part reflects the natural decay of the network if interrelation between nodes ceased. Each node is affected by other nodes. This influence is encoded in a nonlinear term, where the sigmoidal function f [12] depends on a linear combination of all x_i, multiplied by the corresponding factor. The larger this factor is, the more intensive is the influence of x_j to x_i. All information about interrelation between nodes is contained in the so called *regulatory matrix* (sometimes called *weight matrix*)

$$W = \begin{vmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{vmatrix}. \tag{2}$$

The element w_{ij} serves as a measure of the impact of the x_j on x_i. There are other parameters in the system. Constants μ_i are the individual characteristics (gain measure) of any x_i. The constants θ_i serve as thresholds upon reaching which the influence of other nodes begins to be felt. When the rules for all members of the model are known, the future state of the network can be computed. The important question about the system (1) and the corresponding network is obtaining information about attractors. We consider several important cases of regulatory matrices W and thus the respective types of interrelation in a network. For these systems, we provide description of attractors and critical points. Some of them are non-attracting and, taking into account, that the vector field on the boundary of the working region is directed inside the region, the trajectories of the system can tend to attracting sets. The process of pattern formation is also in the focus of our research. The paper contains several sections. First, the technical means are described and needed formulas are presented. Then, the particular cases of the behavior of trajectories are treated such as the activation case, inhibition case, the triangular matrix W case. Finally, several examples are given that show the possibility of all cases and the coefficient conditions for the above-mentioned behaviors are provided. The related results and the reviews can be found in [1–4, 10].

2 Preliminary Results

The nullclines for the system are defined by the relations

$$\begin{cases} x_1 = \dfrac{1}{v_1}\dfrac{1}{1+e^{-\mu_1\,(w_{11}x_1+w_{12}x_2+w_{13}x_3-\theta_1)}}, \\[2ex] x_2 = \dfrac{1}{v_2}\dfrac{1}{1+e^{-\mu_2\,(w_{21}x_1+w_{22}x_2+w_{23}x_3-\theta_2)}}, \\[2ex] x_3 = \dfrac{1}{v_3}\dfrac{1}{1+e^{-\mu_2\,(w_{21}x_1+w_{22}x_2+w_{33}x_3-\theta_3)}}. \end{cases} \tag{3}$$

The critical points for the system (1) are the cross points of the nullclines. They can be found from the system

$$\begin{cases} x_1 - \dfrac{1}{v_1}\dfrac{1}{1+e^{-\mu_1\,(w_{11}x_1+w_{12}x_2+w_{13}x_3-\theta_1)}} = 0, \\[2ex] x_2 - \dfrac{1}{v_2}\dfrac{1}{1+e^{-\mu_2\,(w_{21}x_1+w_{22}x_2+w_{23}x_3-\theta_2)}} = 0, \\[2ex] x_3 - \dfrac{1}{v_3}\dfrac{1}{1+e^{-\mu_2\,(w_{21}x_1+w_{22}x_2+w_{33}x_3-\theta_3)}} = 0. \end{cases} \tag{4}$$

The linearized system for any critical point (x_1^*, x_2^*, x_3^*) is

$$\begin{cases} u_1' = -v_1u_1 + \mu_1w_{11}g_1u_1 + \mu_1w_{12}g_1u_2 + \mu_1w_{13}g_1u_3, \\ u_2' = -v_2u_2 + \mu_2w_{21}g_2u_1 + \mu_2w_{22}g_2u_2 + \mu_2w_{23}g_2u_3, \\ u_3' = -v_3u_3 + \mu_3w_{31}g_3u_1 + \mu_3w_{32}g_3u_2 + \mu_3w_{33}g_3u_3, \end{cases} \tag{5}$$

where

$$g_1 = \frac{e^{-\mu_1(w_{11}x_1^*+w_{12}x_2^*+w_{13}x_3^*-\theta_1)}}{[1+e^{-\mu_1(w_{11}x_1^*+w_{12}x_2^*+w_{13}x_3^*-\theta_1)}]^2}, \tag{6}$$

$$g_2 = \frac{e^{-\mu_2(w_{21}x_1^*+w_{22}x_2^*+w_{23}x_3^*-\theta_2)}}{[1+e^{-\mu_2(w_{21}x_1^*+w_{22}x_2^*+w_{23}x_3^*-\theta_2)}]^2}, \tag{7}$$

$$g_3 = \frac{e^{-\mu_3(w_{31}x_1^*+w_{32}x_2^*+w_{33}x_3^*-\theta_3)}}{[1+e^{-\mu_3(w_{31}x_1^*+w_{32}x_2^*+w_{33}x_3^*-\theta_3)}]^2}. \tag{8}$$

One has

$$A - \lambda I = \begin{vmatrix} \mu_1w_{11}g_1 - v_1 - \lambda & \mu_1w_{12}g_1 & \mu_1w_{13}g_1 \\ \mu_2w_{21}g_2 & \mu_2w_{22}g_2 - v_2 - \lambda & \mu_2w_{23}g_2 \\ \mu_3w_{31}g_3 & \mu_3w_{32}g_3 & \mu_3w_{33}g_3 - v_3 - \lambda \end{vmatrix} \tag{9}$$

and the characteristic equation for $v_1 = v_2 = v_3 = 1$ is

$$\begin{aligned}
&\det|A - \lambda I| = -\Lambda^3 + (\mu_1 w_{11} g_1 + \mu_2 w_{22} g_2 + \mu_3 w_{33} g_3)\Lambda^2 \\
&+[\mu_1\mu_3 g_1 g_3 (w_{31}w_{13} - w_{11}w_{33}) + \mu_2\mu_3 g_2 g_3 (w_{32}w_{23} - w_{22}w_{33}) \\
&+\mu_1\mu_2 g_1 g_2 (w_{21}w_{12} - w_{11}w_{22})]\Lambda \\
&-\mu_1\mu_2\mu_3 g_1 g_2 g_3 (w_{11}w_{32}w_{23} + w_{21}w_{12}w_{33} + w_{31}w_{22}w_{13} \\
&-w_{11}w_{22}w_{33} - w_{12}w_{23}w_{31} - w_{13}w_{21}w_{32}) = 0,
\end{aligned} \tag{10}$$

where $\Lambda = \lambda + 1$.

2.1 All Zeros on the Diagonal of the Regulatory Matrix

Set $w_{11} = w_{22} = w_{33} = 0$. The regulatory matrix is

$$W = \begin{vmatrix} 0 & w_{12} & w_{13} \\ w_{21} & 0 & w_{23} \\ w_{31} & w_{32} & 0 \end{vmatrix} \tag{11}$$

and the system of differential equations takes the form

$$\begin{cases}
x_1' = \dfrac{1}{1 + e^{-\mu_1(w_{12}x_2 + w_{13}x_3 - \theta_1)}} - x_1, \\
x_2' = \dfrac{1}{1 + e^{-\mu_2(w_{21}x_1 + w_{23}x_3 - \theta_2)}} - x_2, \\
x_3' = \dfrac{1}{1 + e^{-\mu_3(w_{31}x_1 + w_{32}x_2 - \theta_3)}} - x_3.
\end{cases} \tag{12}$$

The linearized system for a critical point (x_1^*, x_2^*, x_3^*) is then

$$\begin{cases}
u_1' = -u_1 + \mu_1 w_{12} g_1 u_2 + \mu_1 w_{13} g_1 u_3, \\
u_2' = -u_2 + \mu_2 w_{21} g_2 u_1 + \mu_2 w_{23} g_2 u_3, \\
u_3' = -u_3 + \mu_3 w_{31} g_3 u_1 + \mu_3 w_{32} g_3 u_2,
\end{cases} \tag{13}$$

where g_1, g_2, g_3, given in (6) to (8), are adapted to the case of the regulatory matrix (11). The characteristic equation is

$$-\Lambda^3 + B\Lambda + C = 0, \tag{14}$$

where $\Lambda = \lambda + 1$,

$$B = \mu_1\mu_3 g_1 g_3 (w_{31}w_{13}) + \mu_2\mu_3 g_2 g_3 (w_{32}w_{23}) + \mu_1\mu_2 g_1 g_2 (w_{21}w_{12}), \tag{15}$$

$$C = \mu_1\mu_2\mu_3 g_1 g_2 g_3 (w_{12}w_{23}w_{31} + w_{13}w_{21}w_{32}). \tag{16}$$

For further analysis let us recall the Cardano formulas applied to the equation

$$y^3 + py + q = 0. \tag{17}$$

It has complex roots if

$$Q := \left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2 \tag{18}$$

is positive. The complex roots are given by expressions

$$y_{2,3} = -\frac{a+b}{2} \pm i(a-b)\frac{\sqrt{3}}{2}, \tag{19}$$

where

$$a = (-\frac{q}{2} + \sqrt{Q})^{\frac{1}{3}}, \quad b = (-\frac{q}{2} - \sqrt{Q})^{\frac{1}{3}}$$

are real cubic roots satisfying $a \cdot b = -\frac{p}{3}$ [8, §38]. The remaining real root of equation (17) $y_1 = a + b$ is real.

3 Conditions for a Critical Point to Be a Focus

Consider the case described in Subsect. 2.1. Returning to our notation, we get

$$Q := -\left(\frac{B}{3}\right)^3 + \left(\frac{C}{2}\right)^2. \tag{20}$$

Suppose that $Q > 0$. The characteristic numbers λ for a given critical point (x_1^*, x_2^*, x_3^*) are

$$\begin{aligned} \lambda_1 &= -1 + (a+b), \\ \lambda_{2,3} &= -1 - \frac{a+b}{2} \pm i(a-b)\frac{\sqrt{3}}{2}, \end{aligned} \tag{21}$$

where

$$a = \left(\frac{C}{2} + \sqrt{Q}\right)^{\frac{1}{3}}, \quad b = \left(\frac{C}{2} - \sqrt{Q}\right)^{\frac{1}{3}} \tag{22}$$

are the real values of cubic roots, Q is given by (20). We will call such a critical point 3D-focus. If the real part $-1 - \frac{a+b}{2}$ is positive, this is unstable 3D-focus.

Proposition 3.1. *If $Q > 0$ or, which is the same,*

$$\left(\frac{C}{2}\right)^2 > \left(\frac{B}{3}\right)^3 \tag{23}$$

for a critical point (x_1^*, x_2^*, x_3^*) of the system (12), then this point is a 3D-focus.

Proof. Follows from (20) to (22).

Corollary 1. *If $B < 0$ for some critical point, then this point is a 3D-focus.*

Proof. The relation (23) is fulfilled, if $B < 0$.

Proposition 3.2. *If system (12) has a critical point of type focus then it is unstable focus only if $-1 - \frac{a+b}{2}$, that is, the real part of $\lambda_{2,3}$ in (21), is a positive value.*

Proof. Follows from (21).

4 Inhibition-Activation

Consider the system

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-\mu_1(w_{12}x_2+w_{13}x_3-\theta_1)}} - x_1, \\ x_2' = \dfrac{1}{1+e^{-\mu_2(w_{21}x_1+w_{23}x_3-\theta_2)}} - x_2, \\ x_3' = \dfrac{1}{1+e^{-\mu_3(w_{31}x_1+x_2-\theta_3)}} - x_3, \end{cases} \tag{24}$$

where w_{12}, w_{13}, w_{23} are negative, w_{21}, w_{31}, w_{32} are positive.

We consider the specific case

$$W = \begin{vmatrix} 0 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{vmatrix}, \tag{25}$$

$\mu_1 = \mu_2 = \mu_3 = \mu$, $\theta_1 = \theta_2 = \theta_3 = \theta$. The system then has a single critical point. Introduce

$$g_1 = \frac{e^{-\mu(-x_2-x_3-\theta)}}{[1+e^{-\mu(-x_2-x_3-\theta)}]^2} \tag{26}$$

$$g_2 = \frac{e^{-\mu(\ \ x_1-x_3-\theta)}}{[1+e^{-\mu(\ \ x_1-x_3-\theta)}]^2} \tag{27}$$

$$g_3 = \frac{e^{-\mu(\ \ x_1+x_2-\theta)}}{[1+e^{-\mu(\ \ x_1+x_2-\theta)}]^2} \tag{28}$$

Values of g_i are in the range $(0, 1)$. The linearized system now is

$$\begin{cases} u_1' = -u_1 - \mu g_1 u_2 - \mu g_1 u_3, \\ u_2' = \mu g_2 u_1 - u_2 - \mu g_2 u_3, \\ u_3' = \mu g_3 u_1 + \mu g_3 u_2 - u_3. \end{cases} \tag{29}$$

The characteristic equation can be obtained from

$$A - \lambda I = \begin{vmatrix} -1-\lambda & -\mu g_1 & -\mu g_1 \\ \mu g_2 & -1-\lambda & -\mu g_2 \\ \mu g_3 & \mu g_3 & -1-\lambda \end{vmatrix} \tag{30}$$

and

$$\det|A - \lambda I| = -\lambda^3 - 3\lambda^2 - \mu^2(g_1g_2 + g_1g_3 + g_2g_3)(\lambda + 1) - 3\lambda - 1 = 0. \tag{31}$$

The characteristic numbers are

$$\begin{cases} \lambda_1 = -1, \\ \lambda_2 = -1 - \mu\sqrt{g_1g_2 + g_1g_3 + g_2g_3}\, i, \\ \lambda_3 = -1 + \mu\sqrt{g_1g_2 + g_1g_3 + g_2g_3}\, i. \end{cases} \tag{32}$$

Proposition 4.1. *A critical point of the system (24) under the above conditions is 3D-focus, that is, the following is true: there is 2D-subspace with a stable focus and attraction in the remaining dimension.*

5 Triangular System

We consider the specific case of the regulatory matrix

$$W = \begin{vmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ 0 & w_{22} & \dots & w_{2n} \\ \dots & & & \\ 0 & 0 & \dots & w_{nn} \end{vmatrix}, \tag{33}$$

but in the n-dimensional variant. The system of differential equations takes the form

$$\begin{cases} x_1' = \dfrac{1}{1 + e^{-\mu_1(w_{11}x_1 + w_{12}x_2 + \dots + w_{1n}x_n - \theta_1)}} - x_1, \\ x_2' = \dfrac{1}{1 + e^{-\mu_2(\quad w_{22}x_2 + \dots + w_{2n}x_n - \theta_2)}} - x_2, \\ \dots \\ x_n' = \dfrac{1}{1 + e^{-\mu_n(\qquad w_{nn}x_n - \theta_n)}} - x_n, \end{cases} \tag{34}$$

where $n > 1$. Constants w_{ij} take values in (0; 1].

5.1 Critical Points

Critical points of the system (34) are to be determined from

$$\begin{cases} x_1 = \dfrac{1}{1+e^{-\mu_1(w_{11}x_1+w_{12}x_2+\ldots+w_{1n}x_n-\theta_1)}}, \\ x_2 = \dfrac{1}{1+e^{-\mu_2(\qquad w_{22}x_2+\ldots+w_{2n}x_n-\theta_2)}}, \\ \ldots \\ x_n = \dfrac{1}{1+e^{-\mu_n(\qquad\qquad w_{nn}x_n-\theta_n)}}. \end{cases} \tag{35}$$

Since the right sides in (35) are positive but less than unity, all critical points locate in (0; 1) × (0; 1) × ... × (0; 1).

We claim that there are only three possibilities for the number of critical points for the x_n in the system (34).

Proposition 5.1. *There are at most three values for x_n in the system (35).*

Proposition 5.2. *The system (34) has at most 3^n critical points.*

Proof. Consider the two last equations of the system (35). The last one has at most three critical points. This is true due to the S-shape graph of the sigmoidal function in the right side of the last equation in (35). Putting each of them into the penultimate equation and taking into account that for any x_n it can have at most three roots x_{n-1}, at most nine values of x_{n-1} can be obtained. Proceeding in this manner, we obtain at most 3^n critical points.

5.2 Linearized System

The linearized system is

$$\begin{cases} u_1' = -u_1 + \mu_1 w_{11} g_1 u_1 + \mu_1 w_{12} g_1 u_2 + \ldots + \mu_1 w_{1n} g_1 u_n, \\ u_2' = -u_2 + \mu_2 w_{22} g_2 u_2 + \ldots + \mu_2 w_{2n} g_2 u_n, \\ \ldots \\ u_n' = -u_n + \mu_n w_{nn} g_n u_n, \end{cases} \tag{36}$$

where

$$g_1 = \frac{\mu_1 e^{-\mu_1(w_{11}x_1+w_{12}x_2+\ldots+w_{1n}x_n-\theta_1)}}{[1+e^{-\mu_1(w_{11}x_1+w_{12}x_2+\ldots+w_{1n}x_n-\theta_1)}]^2} \tag{37}$$

$$g_2 = \frac{\mu_2 e^{-\mu_2(w_{22}x_2+\ldots+w_{2n}x_n-\theta_2)}}{[1+e^{-\mu_2(w_{22}x_2+\ldots+w_{2n}x_n-\theta_2)}]^2} \tag{38}$$

...

$$g_n = \frac{\mu_n e^{-\mu_n(w_{nn}x_n-\theta_n)}}{[1+e^{-\mu_n(w_{nn}x_n-\theta_n)}]^2}. \tag{39}$$

Values of g_i are positive and less than unity. The characteristic values for a critical point are to be obtained from

$$A-\lambda I = \begin{vmatrix} \mu_1 w_{11} g_1 - 1 - \lambda & \mu_1 w_{12} g_1 & \dots & \mu_1 w_{1n} g_1 \\ 0 & \mu_2 w_{22} g_2 - 1 - \lambda & \dots & \mu_2 w_{2n} g_2 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mu_n w_{nn} g_n - 1 - \lambda \end{vmatrix} \tag{40}$$

and

$$\begin{array}{l} \det|A-\lambda I| = (\mu_1 w_{11} g_1 - 1 - \lambda)(\mu_2 w_{22} g_2 - 1 - \lambda)\dots \\ \dots(\mu_n w_{nn} g_n - 1 - \lambda) = 0. \end{array}. \tag{41}$$

Evidently

$$\begin{cases} \lambda_1 = -1 + \mu_1 w_{11} g_1, \\ \lambda_2 = -1 + \mu_2 w_{22} g_2, \\ \dots \\ \lambda_n = -1 + \mu_n w_{nn} g_n. \end{cases} \tag{42}$$

If $\mu_i w_{ii} g_i > 1$ then $\lambda_i > 0$.
If $\mu_i w_{ii} g_i = 1$ then $\lambda_i = 0$.
If $\mu_i w_{ii} g_i < 1$ then $\lambda_i < 0$.

Proposition 5.3. *The system (34) cannot have critical points of type focus.*

Examples

Example 1. For the system with the triangular matrix W

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-3(x_1+x_2+x_3-1)}} - x_1, \\ x_2' = \dfrac{1}{1+e^{-3(x_2+x_3-1)}} - x_2, \\ x_3' = \dfrac{1}{1+e^{-3(x_3-1)}} - x_3, \end{cases} \tag{43}$$

the unique critical point is $(0.084922, 0.0671045, 0.0555481)$. The values of λ for this critical point are

$$\begin{cases} \lambda_1 = -1.23313, \\ \lambda_2 = -1.1878, \\ \lambda_3 = -1.15739. \end{cases} \tag{44}$$

Example 2. Consider the system given in [9]

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-7(x_2-6.2x_3-0.5)}} - 0.62x_1, \\ x_2' = \dfrac{1}{1+e^{-7(x_1-0.3)}} - 0.42x_2, \\ x_3' = \dfrac{1}{1+e^{-13(x_1-0.7)}} - 0.1x_3. \end{cases} \tag{45}$$

It has a single critical point at $(0.4015, 1.5966, 0.2023)$. The characteristic values for this point are $\lambda_1 = -1.8306$, $\lambda_{2,3} = 0.345301 \pm 0.139015i$. This critical point is not attractive due to the positive real parts of $\lambda_{2,3}$.

Example 3. The system

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-7(x_2-6.2x_3-0.5)}} - 0.62x_1, \\ x_2' = \dfrac{1}{1+e^{-7(x_1-x_3-0.3)}} - 0.42x_2, \\ x_3' = \dfrac{1}{1+e^{-13(x_1+x_2-0.7)}} - 0.1x_3 \end{cases} \tag{46}$$

has a single critical point at $(0.037648, 0.247582, 0.0453218)$. The characteristic values for this point are $\lambda_1 = -0.860672$, $\lambda_{2,3} = -0.139664 \pm 0.0780681i$. This critical point is attractive.

Example 4. Further modification of the Example 2 leads to the system

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-7(x_2-6.2x_3-0.5)}} - 0.62x_1, \\ x_2' = \dfrac{1}{1+e^{-7(x_1-x_3-0.3)}} - 0.42x_2, \\ x_3' = \dfrac{1}{1+e^{-13(x_1+0.1x_2-0.7)}} - 0.1x_3, \end{cases} \tag{47}$$

which has a single critical point at $(0.251009, 0.710983, 0.072996)$. The characteristic values for this point are $\lambda_1 = -1.55538$, $\lambda_{2,3} = 0.207688 \pm 0.412065i$. This critical point is not attractive. There is a periodic solution that can be drawn using the initial values $(0.032, 0.239, 0.03)$.

Example 5. Removing the coefficient 6.2 in the first equation of the system (47) yields

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-7(x_2-x_3-0.5)}} - 0.62x_1, \\ x_2' = \dfrac{1}{1+e^{-7(x_1-x_3-0.3)}} - 0.42x_2, \\ x_3' = \dfrac{1}{1+e^{-13(x_1+0.1x_2-0.7)}} - 0.1x_3. \end{cases} \tag{48}$$

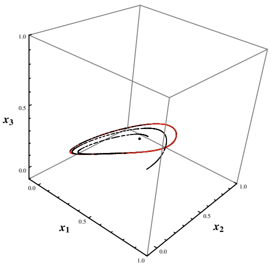


Fig. 1 The periodic solution in Example 4

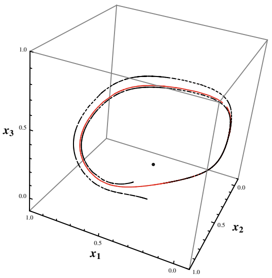


Fig. 2 The periodic solution in Example 5

This system has a single critical point at (0.356279, 0.548522, 0.228589). The characteristic equation is

$$-\lambda^3 - 1.14\lambda^2 + 0.570414\lambda - 0.490819 = 0.$$

The characteristic values are $\lambda_1 = -1.66122$, $\lambda_{2,3} = 0.260612 \pm 0.477009i$. This critical point is not attractive. The periodic solution can be drawn using the initial values (0.064, 0.2, 0.08) (Figs. 1 and 2).

Example 6. The system

$$\begin{cases} x_1' = \dfrac{1}{1+e^{-7(x_2-x_3-0.5)}} - 0.52x_1, \\ x_2' = \dfrac{1}{1+e^{-7(x_1-x_3-0.3)}} - 0.42x_2, \\ x_3' = \dfrac{1}{1+e^{-13(x_1+0.1x_2-0.7)}} - 0.1x_3 \end{cases} \tag{49}$$

has a single critical point at (0.36187, 0.53125, 0.240095). The characteristic values are $\lambda_1 = -1.61197$, $\lambda_{2,3} = 0.285986 \pm 0.466687i$. This critical point is not attractive. The three trajectories tend to a periodic solution depicted in Fig. 3.

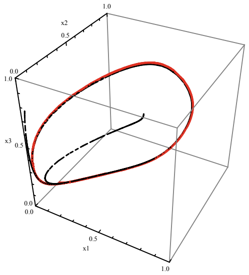


Fig. 3 The periodic solution in Example 6

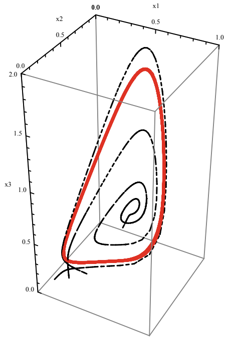


Fig. 4 The periodic solution in Example 7

Example 7. The system

$$
\begin{cases}
x_1' = \dfrac{1}{1+e^{-10(x_2-x_3-0.5)}} - 0.09x_1, \\
x_2' = \dfrac{1}{1+e^{-5(x_1-0.1x_3-0.3)}} - 0.9x_2, \\
x_3' = \dfrac{1}{1+e^{-20(x_1+0.1x_2-0.7)}} - 0.09x_3
\end{cases}
\tag{50}
$$

has a single critical point at $(0.478312, 0.723644, 0.533787)$. The characteristic equation is

$$-\lambda^3 - 1.08\lambda^2 - 0.08948\lambda - 0.390813 = 0.$$

The characteristic values are $\lambda_1 = -1.25637$, $\lambda_{2,3} = 0.0881848 \pm 0.550717i$. This critical point is not attractive. The trajectories tend to a periodic solution depicted in Fig. 4.

The initial conditions for the periodic solutions in Examples 4 to 7 were found by observing the projections of solutions on the three coordinate planes. Since any nearby solution tends to the periodic one, starting from some moment all three projections become periodic.

6 Conclusions

The behavior of solutions of systems of the form (1) strongly depends on the structure of the weight matrix W. Any system (1) has at least one critical point in the region $D = (0, 1/v_1) \times (0, 1/v_2) \times (0, 1/v_3)$. No trajectory of the system (1) can escape this region. Multiple critical points are possible. Stable nodes, stable and unstable 3D-focuses and saddle points can occur. Systems with a triangular matrix W cannot have focuses. Inhibition-activation systems of Sect. 4 have a critical point that is a focus. The coefficient conditions are possible for a critical point to be a focus. No attracting critical points may exist in D. The trajectories tend then to a pattern of regular form. In the examples 4 to 7 periodic solutions are detected numerically and have been visualized together with some neighboring trajectories. No chaotic behavior was observed yet.

References

1. Alakwaa, F.: Modeling of gene regulatory networks: a literature review. J. Comp. Syst. Biol. **1**(1), 1–8 (2014)
2. Brokan, E., Sadyrbaev, F.: On a differential system arising in the network control theory (2016)
3. Brokan, E., Sadyrbaev, F.: On attractors in gene regulatory systems. In: Fořt, J., Fürst, J., Halama, J., Herbin, R., Hubert, F. (eds.) Finite Volumes for Complex Applications VI Problems & Perspectives, AIP Conference Proceedings, vol. 1809, pp. 020,010–1–020,010–9. AIP Publishing (2017)
4. Brokan, E., Sadyrbaev, F.: Attraction in n-dimensional differential systems from network regulation theory. Math. Methods Appl. Sci. **41**(17), 7498–7509 (2018)
5. Furusawa, C., Kaneko, K.: A generic mechanism for adaptive growth rate regulation. PLoS Comput. Biol. **4**(1), 0035–0042 (2008)
6. Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. **9**(1), 67–103 (2002)
7. Koizumi, Y.: Adaptive virtual network topology control based on attractor selection. J. Lightwave Technol. **28**(11), 1720–1731 (2010). iSSN 0733-8724
8. Kurosh, A.: Higher Algebra. MIR Publishers, Moscow (1972). (in Russian)
9. Mukherjee, S.: Is one dimensional poincare map sufficient to describe the chaotic dynamics of a three dimensional system? Appl. Math. Comput. **219**(23), 11,056–11,064 (2013)
10. Vijesh, N.: Modeling of gene regulatory networks: a review. J. Biomed. Eng. **6**(2A), 223–231 (2013)
11. Vohradský, J.: Neural network model of gene expression. FASEB J. **15**(3), 846–854 (2001)
12. wiki: Sigmoid function. Technical report (2020). https://en.wikipedia.org/wiki/Sigmoid_function

Analytic Traveling-Wave Solutions of the Kardar-Parisi-Zhang Interface Growing Equation with Different Kind of Noise Terms

I. F. Barna, G. Bognár, L. Mátyás, M. Guedda, and K. Hriczó

Abstract The one-dimensional Kardar-Parisi-Zhang dynamic interface growth equation with the traveling-wave Ansatz is analyzed. As a new feature additional analytic terms are added. From the mathematical point of view, these can be considered as various noise distribution functions. Six different cases were investigated among others Gaussian, Lorentzian, white or even pink noise. Analytic solutions are evaluated and analyzed for all cases. All results are expressible with various special functions Mathieu, Bessel, Airy or Whittaker functions showing a very rich mathematical structure with some common general characteristics. This study is the continuation of our former work, where the same physical phenomena was investigated with the self-similar Ansatz. The differences and similarities among the various solutions are enlightened.

I. F. Barna
Wigner Research Center, P.O. Box 49, Budapest 1525, Hungary
e-mail: barna.imre@wigner.mta.hu

G. Bognár (✉)
Faculty of Mechanical Engineering and Informatics, University of Miskolc,
Miskolc-Egyetemváros 3515, Hungary
e-mail: v.bognar.gabriella@uni-miskolc.hu

L. Mátyás
Faculty of Economics, Socio-Human Sciences and Engineering, Department of Bioengineering,
Sapientia Hungarian University of Transylvania, Libertătii sq. 1, 530104 Miercurea Ciuc, Romania
e-mail: matyaslaszlo@uni.sapientia.ro

M. Guedda
Faculte de Mathematiques et d'Informatique, Université de Picardie Jules Verne Amiens,
33, rue Saint-Leu, 80039 Amiens, France
e-mail: mohamed.guedda@u-picardie.fr

K. Hriczó
Faculty of Mechanical Engineering and Informatics, University of Miskolc,
Miskolc-Egyetemváros 3515, Hungary
e-mail: mathk@uni-miskolc.hu

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_19

1 Introduction

Solidification fronts or crystal growth is a scientific topic which attracts much interest from a long time. Basic physics of growing crystallines can be found in large number of textbooks (see e.g., [1]). One of the simplest nonlinear generalization of the ubiquitous diffusion equation is the so called Kardar-Parisi-Zhang (KPZ) model obtained from Langevin equation

$$\frac{\partial u}{\partial t} = \nu \nabla^2 u + \frac{\lambda}{2}(\nabla u)^2 + \eta(\mathbf{x}, t), \tag{1}$$

where u stands for the profile of the local growth [2]. The first term on the right hand side describes relaxation of the interface by a surface tension preferring a smooth surface. The next term is the lowest-order nonlinear term that can appear in the surface growth equation justified with the Eden model. The origin of this term lies in non-equilibrium. The third term is a Langevin noise which mimics the stochastic nature of any growth process and usually has a Gaussian distribution. In the last two decades numerous studies came to light about the KPZ equation. Without completeness we mention some of them. The basic physical background of surface growth can be found in the book of Barabási and Stanley [3]. Later, Hwa and Frey [4, 5] investigated the KPZ model with the help of the renormalization group-theory and the self-coupling method which is a precise and sophisticated method using Green's functions. Various dynamical scaling forms of $C(x, t) = x^{-2\varphi}C(bx, b^z t)$ were considered for the correlation function (where φ, b and z are real constants). The field theoretical approach by Lässig was to derive and investigate the KPZ equation [6]. Kriecherbauer and Krug wrote a review paper [7], where the KPZ equation was derived from hydrodynamical equations using a general current density relation.

Several models exist and all lead to similar equations as the KPZ model, one of them is the interface growth of bacterial colonies [8]. Additional general interface growing models were developed based on the so-called Kuramoto-Sivashinsky (KS) equation which shows similarity to the KPZ model with an extra $\nabla^4 u$ term [9].

Kersner and Vicsek investigated the traveling wave dynamics of the singular interface equation [10] which is closely related to the KPZ equation. One may find certain kind of analytic solutions to the problem [11] as already mentioned in [12].

Ódor and co-worker intensively examined the two dimensional KPZ equation with dynamical simulations to investigate the aging properties of polymers or glasses [13].

Beyond these continuous models based on partial differential equations (PDEs), there are large number of purely numerical methods available to study diverse surface growth effects. As a view we mention the kinetic Monte Carlo [14] model, Lattice-Boltzmann simulations [15], and the etching model [16].

In this paper we investigate the solutions to the KPZ equation with the traveling wave Ansatz in one-dimension applying various forms of the noise term. The effects of the parameters involved in the problem are examined.

2 Theory

In general, non-linear PDEs has no general mathematical theory which could help us to understand general features or to derive physically relevant solutions. Basically, there are two different trial functions (or Ansatz) which have well-founded physical interpretation. The first one is the traveling wave solution, which mimics the wave property of the investigated phenomena described by the non-linear PDE of the form

$$u(x,t) = f(x \pm ct) = f(\omega), \tag{2}$$

where c means the velocity of the corresponding wave. Gliding and Kersner used the traveling wave Ansatz to investigate study numerous reaction-diffusion equation systems [17]. To describe pattern formation phenomena [18] the traveling waves Ansatz is a useful tool as well. Saarloos investigated the front propagation into unstable states [19], where traveling waves play a key role.

This simple trial function can be generalized in numerous ways, e.g., to $e^{-\alpha t} f(x \pm ct) := e^{-\alpha t} f(\omega)$ which describes exponential decay or to $g(t) \cdot f(x \pm c \cdot t) := g(t) f(\omega)$ which can even be a power law function of the time as well. We note, that the application of these Ansatz to the KPZ equation leads to the triviality of $e^{-\alpha t} = g(t) \equiv 1$. In 2006, He and Wu developed the so-called exp-function method [20] which relying on an Ansatz (a rational combination of exponential functions), involving many unknown parameters to be specified at the stage of solving the problem. The method soon drew the attention of many researchers, who described it as “straightforward”, “reliable”, and “effective”. Later, Aslan and Marinakis [21] summarized various applications of the Ansatz.

There is another existing remarkable Ansatz interpolating the traveling-wave and the self-similar Ansatz by Benhamidouche [22].

The second one is the self-similar Ansatz [23] of the form $u(x,t) = t^{-\alpha} f\left(\frac{x}{t^{\beta}}\right) := t^{-\alpha} f(\omega)$. The associated mathematical and physical properties were exhaustively discussed in our former publications [24, 25] or in a book chapter [26] in the field of hydrodynamics. All these kind of methods belong to the so-called reduction mechanism, where applying a suitable variable transformation the original PDEs or systems of PDEs are reduced to an ordinary differential equation (ODE) or systems of ODEs.

3 Results Without the Noise Term

Applying the traveling wave Ansatz to the KPZ PDE with $\eta(x,t) = 0$, Eq. (1) leads to the ODE of

$$-\nu f''(\omega) + f'(\omega)\left[c - \frac{\lambda}{2} f'(\omega)\right] = 0, \tag{3}$$

From now on we use the Maple 12 mathematical program package to obtain analytic solutions in closed forms. For Eq. (3), it can be given as

$$f(\omega) = \frac{2}{\lambda} \ln \left(\frac{\lambda \left[c_1 \nu e^{\frac{c\eta}{\nu}} + c_2 c \right]}{2\nu c} \right) \nu, \tag{4}$$

where c_1 and c_2 are the constants of integration and c is the speed of the wave.

We fix this notation from now on throughout the paper. Note, that this is an equation of a linear function $f(\omega) = a\omega + b$ (just given in a complicated form) with any kind of parameter set, except $c_1 = 0$ which gives a constant solution. This physically means that there is a continuous surface growing till infinity which is quite unphysical. Therefore, some additional noise is needed to have surface growing phenomena. We remark the general properties of all the forthcoming solutions. Due to the Hopf-Cole transformation [27, 28] ($h = A\ln(y)$) convertes the non-linear KPZ equation to the regular heat conduction (or diffusion) equation with an additional stochastic source term eliminating the non-linear gradient-squared term. All the solutions contain a logarithmic function with a complicated argument. In this sense, the solutions have the same structure, the only basic difference is the kind of special function in the argument. If these argument functions take periodically positive and negative real values then the logarithmic function creates distinct intervals (small islands which describe the surface growing mechanisms, and define the final solution). This statement is generally true for our former study as well [29, 30].

Remark that the solution to (1) obtained from the self-similar Ansatz reads

$$f(\omega) = \frac{2\nu}{\lambda} ln \left(\frac{\lambda c_1 \sqrt{\pi\nu}\ \ erf[\omega/(2\sqrt{\nu})] + c_2}{2\nu} \right), \tag{5}$$

where $erf[\ ]$ means the error function [31]. Figure 1 compares these two solutions. We note the asymptotic convergence of the self-similar solution and the divergence of the traveling-wave solution. We have the same conclusion as in our former study [29] (where the self-similar Ansatz was applied), that without any noise term the KPZ equation cannot be applied to describe surface growth phenomena. The different kind of noise terms define different kind of extra islands (parts of the solution having compact supports) and these islands show a growth dynamics.

To have a better understanding between the two solutions, Fig. 2 shows the projection of both complete solutions $u(x, y = 0, t)$. The major differences are still present.

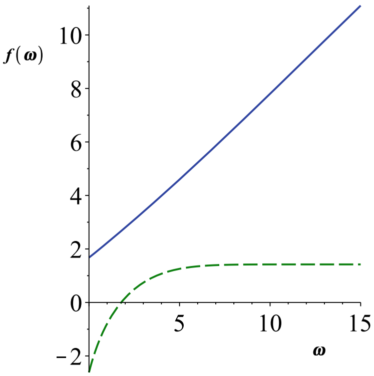


Fig. 1 The two shape functions of the KPZ equation without any kind of noise term. The solid line represents the solution for traveling-wave and the dashed line is for the self-similar Ansatz. The applied parameter set is $c_1 = c_2 = c = 1, \nu = 4, \lambda = 3$

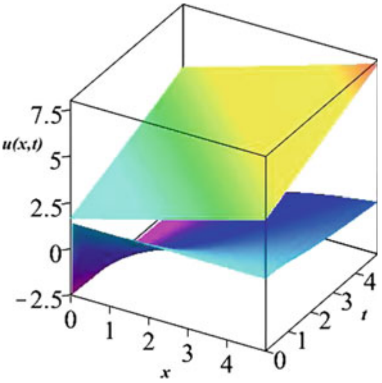


Fig. 2 The two solutions of the KPZ equation without any noise term. The upper lying function represents the traveling-wave solution. The applied parameter set is the same as used above

4 Results with Various Noise Terms

As we mentioned in our former study [29] only the additional noise term makes the KPZ solutions interesting. We search the solutions with the traveling-wave Ansatz, therefore is it necessary that the noise term η should be an analytic function of $\omega = x + ct$ like $\eta(\omega) = a(x + ct)^2$. We will see that for some kind of noise terms it is not possible to find a closed analytic solution when all the physical parameters are free (ν, λ, c, a), however, if some parameters are fixed it becomes possible to find analytic expressions. It is also clear, that it is impossible to perform a mathematically rigorous complete function analysis according to all four physical and two integral parameters c_1, c_2. We performed numerous parameter studies and gave the most relevant parameter dependencies of the solutions.

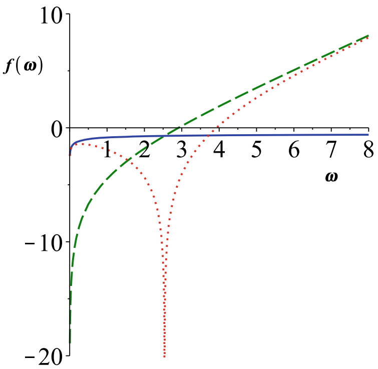


Fig. 3 Three different shape functions for the brown noise $n = -2$. The applied physical parameter set is $\lambda = 5, \nu = 3, a = 2$ and $c = 2$. The dashed line is for $c_1 = 1, c_2 = 0$, the dotted line is for $c_1 = c_2 = 1$ and the solid line is for $c_1 = 0, c_2 = 1$, respectively

4.1 Brown Noise $n = -2$

As first, case let us consider the brown noise $\eta(x,t) = \frac{a}{\omega^2}$. It leads to the following ODE

$$-\nu f''(\omega) + f'(\omega)\left[c - \frac{\lambda}{2} f'(\omega)\right] - \frac{a}{\omega^2} = 0. \tag{6}$$

The solution can be given in the form

$$f(\omega) = \frac{1}{\lambda}\left(c\eta + \nu \ln\left\{\frac{\lambda^2\left[-c_1 I_d\left(\frac{c\omega}{2\nu}\right) + c_2 K_d\left(\frac{c\omega}{2\nu}\right)\right]^2}{c^2\omega\left[K_d\left(\frac{c\omega}{2\nu}\right) I_{d+1}\left(\frac{c\omega}{2\nu}\right) + I_d\left(\frac{c\omega}{2\nu}\right) K_{d+1}\left(\frac{c\omega}{2\nu}\right)\right]^2}\right\}\right) \tag{7}$$

where $I_d(\omega)$ and $K_d(\omega)$ are the modified Bessel functions of the first and second kind [31] with the subscript of $d = \frac{\sqrt{\nu^2 - 2a\lambda}}{2\nu} + 1$. To obtain real solutions for the KPZ equation (which provides the height of the surface) the order of the Bessel function (notated as the subscript) has to be non-negative and provides the following constrain $\nu^2 \geq 2a\lambda$. This gives us a reasonable relation among the three terms of the right hand side of Eq. (1). When the magnitude of the noise term a becomes large enough no surface growth take place. Figure 3 presents solutions with different combinations of the integration constants c_1, c_2. Having in mind, that the $K_d()$ Bessel function of the second kind is regular at infinity, one gets that it has a strong decay at large argument ω. The $c_1 = 0, c_2 = 0$ type solutions have physical relevance. Figure 4 shows the complete solution of the KPZ equation. It can be seen that a sharp and localized peak exists for a short time. Therefore, no typical surface growth phenomena is described with this kind of noise and initial conditions.

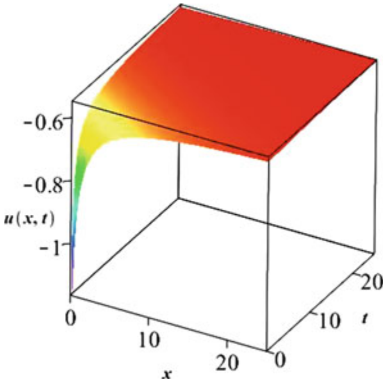


Fig. 4 The solution $u(x,t)$ to the KPZ equation for the brown noise $n=-2$ with the parameter set of $c_1=c_2=c=1, \nu=4, \lambda=3$

4.2 *Pink Noise n = −1*

The noise term $\eta=\frac{a}{\omega}$ corresponds to the ODE

$$-\nu f''(\omega)+f'(\omega)\left[c-\frac{\lambda}{2}f'(\omega)\right]-\frac{a}{\omega}=0, \tag{8}$$

whose general solution is

$$f(\omega)=\frac{1}{\lambda}+\ln\left\{\frac{-c\lambda[c_1M(\varepsilon_b)-c_2U(\varepsilon_b)]}{M(\varepsilon_d)(2\nu cU(\varepsilon_b)+a\lambda U(\varepsilon_b))+2M(\varepsilon_b)\nu cU(\varepsilon_d)}\right\}, \tag{9}$$

where $M(\varepsilon_b)$ and $U(\varepsilon_d)$ are the Kummer M and Kummer U functions (for more see [31]) with the parameters of $\varepsilon_b=(\frac{2c\nu-a\lambda}{2c\nu},2,\frac{c\omega}{\nu})$ and $\varepsilon_d=(\frac{-a\lambda}{2c\nu},2,\frac{c\omega}{\nu})$. Figure 5 shows three different shape functions corresponding to the pink noise. The evaluation of direct parameter dependencies of the solutions are not trivial. In some reasonable parameter range we found the following trends: for fixed a, c, ν and larger λ values, the solution shows more independent well-defined "bumps" or islands and higher steepness of the line which connects the maxima of the existing peaks of the islands. At fixed parameter values a, c, λ, different values of ν just shift the position of the existing peaks. The role of a and c is not defined. Figure 6 presents a total solution $u(x,t)$ to the KPZ equation, the freely traveling three islands are clearly seen.

4.3 *White Noise n = 0*

Here, the noise term is $\eta=a\omega^0=a$ which leads to the ODE of

Fig. 5 Three different shape functions for the pink noise ($n = -1$). Solid, dashed and dotted lines are for the parameter sets of ($c_1 = c_2 = 1; c = 1/2, \nu = 0.85, \lambda = 3, a = 2$), ($c_1 = c_2 = 1; c = 1/2, \nu = 0.85, \lambda = 2.5, a = 2$), ($c_1 = c_2 = 1; c = 0.6, \nu = 0.85, \lambda = 5, a = 2$), respectively

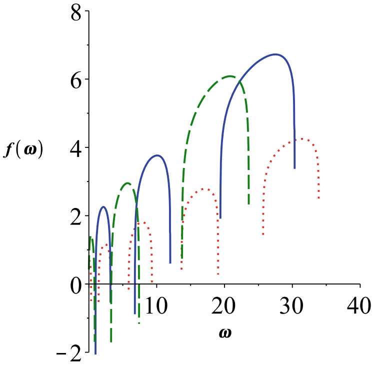
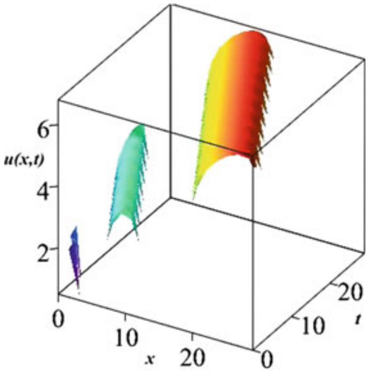


Fig. 6 The total solution of the KPZ equation for $n = -1$ with the applied parameter set $c_1 = c_2 = 1, c = 1/2, \nu = 0.85\lambda = 3$ and $a = 2$



$$- \nu f''(\omega) + f'(\omega)\left[c - \frac{\lambda}{2} f'(\omega)\right] - a = 0, \tag{10}$$

$$f(\omega) = \frac{\omega c}{\lambda} - \frac{\omega\sqrt{c^2 - 2a\lambda}}{\lambda} - \frac{2\nu \ln(2)}{\lambda} - \frac{\nu \ln\left(\frac{c^2-2a\lambda}{\lambda^2\left[c_1 e^{\frac{\omega\sqrt{c^2-2a\lambda}}{\nu}} - c_2\right]^2}\right)}{\lambda} \tag{11}$$

Figure 7 shows two shape functions for two different parameter sets. There exists basically two different functions depending on the ratios of the integral constants c_1 and c_2. The first is a pure linear function with infinite range and its domain represents boundless surface growth, which is a physical nonsense. The second solution is a sum of a linear and logarithmic function with a domain bounded from above due to the argument of the ln function. Figure 8 shows the final solution of the KPZ equation $u(x, t)$. We note that with the substitution $\omega = x + ct$ only the first kind of solution

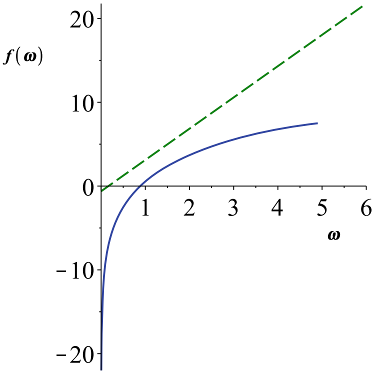


Fig. 7 Two shape functions for the constant or white noise. The solid line is for the parameter set $c_1 = 4, c_2 = -1, c = 0.3, \nu = 2, \lambda = 1, a = 1$, and the dashed line is for $c_1 = c_2 = 1, c = 4, \nu = 0.5, \lambda = 1, a = 0.3$, respectively

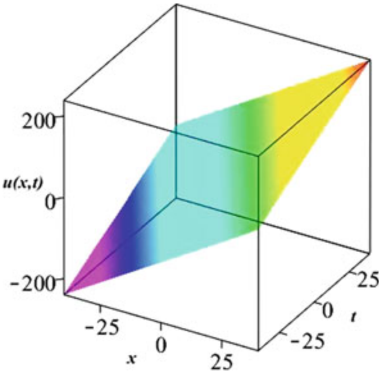


Fig. 8 The KPZ solution for the constant or white noise. The applied parameter set is $c_1 = c_2 = 1, c = 4, \nu = 0.5, \lambda = 1, a = 0.3$, respectively

remains real. For the second parameter set which creates a modified ln function with a cut at well-defined argument becomes complex.

4.4 Blue Noise n = 1

The last color noise $\eta = a\omega$ leads to the ODE of

$$-\nu f''(\omega) + f'(\omega)\left[c - \frac{\lambda}{2}f'(\omega)\right] - a\omega = 0, \tag{12}$$

with the general solution of

$$f(\omega) = \frac{c\omega}{\lambda} - \frac{4\nu \ln(2)}{3\lambda} + \frac{2\nu}{3\lambda}\ln\left\{\frac{\lambda^2[c_1 Ai(\tilde{\omega}) - c_2 Bi(\tilde{\omega})]^3}{\nu a[Ai(1,\tilde{\omega})Bi(1,\tilde{\omega}) - Bi(1,\tilde{\omega})Ai(\tilde{\omega}))]^3}\right\} \tag{13}$$

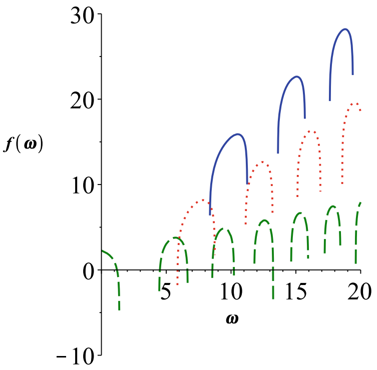


Fig. 9 The shape function for the blue noise for three parameter sets. The solid, dashed and dotted lines are for the parameter sets ($c_1 = 1, c_2 = 0, c = 3, a = 0.5, \nu = 1.5, \lambda = 2$), ($c_1 = c_2 = c = 1, a = 1, \nu = 1, \lambda = 3$), and ($c_1 = c_2 = c = 1, a = 1, \nu = 1, \lambda = 3$), respectively

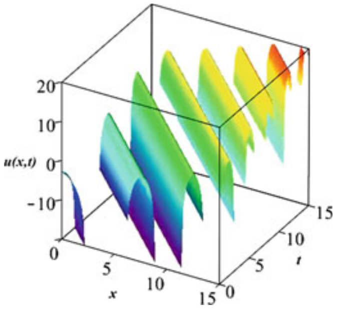


Fig. 10 The solution $u(x, t)$ for the $n = 1$ or blue noise with the applied parameter set of $c_1 = c_2 = c = 1, \nu = 2, \lambda = 3, a = 1$

where $Ai(\tilde{\omega})$, $Bi\tilde{\omega})$ denote the Airy functions of the first and second kind and $Ai(1, \tilde{\omega})$ and $Bi(1, \tilde{\omega})$ are the first derivatives of the Airy functions, where we used the following notation: $\tilde{\omega} = \frac{-(2a\omega\lambda - c^2)4^{\frac{1}{3}}\left(\frac{a\lambda}{\nu^2}\right)^{1/3}}{4a\lambda}$. Exhaustive details of the Airy function can be found in [32]. When the argument ω is positive, $Ai(\omega)$ is positive, convex, and decreasing exponentially to zero, while $Bi(\omega)$ is positive, convex, and increasing exponentially. When ω is negative, $Ai\omega)$ and $Bi\omega)$ oscillate around zero with ever-increasing frequency and ever-decreasing amplitude.

Figure 9 represents shape functions with different parameter sets. Our analysis showed that the composite argument of the ln function is purely real having a decaying oscillatory behavior with alternatively positive and negative values. The ln function creates infinite number of separate "bumps" or islands with compact supports and infinite first spatial derivatives at their boarders. Combining the first two terms of the (13), we get an infinite series of separate islands with increasing height. The ratio c/λ is the steepness of the line, this automatically defines the steepness of the absolute height of the islands. The effects of the various parameters are not quite independent and hard to define, we may say that in general each parameter ν, λ, a, c alone can change the widths, spacing and absolute height of the peaks. Figure 10 shows the total solution of the KPZ equation. The traveling "bumps" are clearly visible.

4.5 Lorentzian Noise

As a first non-colour noise let us consider the Lorentzian noise of the form $\eta = \frac{a}{1+\omega^2}$. It leads to the ODE of

$$-\nu f''(\omega) + f'(\omega)\left[c - \frac{\lambda}{2} f'(\omega)\right] - \frac{a}{1+\omega^2} = 0, \tag{14}$$

We mention, that for the classical exponential and Gaussian noise distributions we could not give solutions in closed analytic form. Unfortunately, there is no closed analytic expression available if all the parameters (ν, λ, c, a) are free. The formal solution contains integrals of the Heun C confluent functions multiplied by some polynomials. However, if the parameters a, λ, ν are fixed, there is analytic solution available for free propagation speed c. The exact solution for $a = \lambda = \nu = 1/2$, and $c = 2$ is the following

$$f(\omega) = c\omega + 2\ln\left\{\frac{c_1 C(B) - c_2 \omega C(A)}{2(\omega^4+\omega^2)[C(A)C'(B) - C(B)C'(A)] + (1+\omega^2)C(A)C(B)}\right\}, \tag{15}$$

where $C'()$ means the first derivative of the Heun C function [33]. For the better transparency we introduce the following notations $A = 0, \frac{1}{2}, 1, \frac{c^2}{4}, 1 - \frac{c^2}{4}; -\omega^2$ and $B = 0, -\frac{1}{2}, 1, \frac{c^2}{4}, 1 - \frac{c^2}{4}; -\omega^2$.

Figure 11 shows the shape function for given parameter set. There is a broad island close to the origin and numerous tiny ones at larger arguments. The numerical accuracy of Maple 12 was enhanced to reach this resolution. It is well-known that the Heun functions are the most complicated objects among special functions and the evaluations needs more computer time.

Figure 12 presents the total solution of the original KPZ. Due to the substitution $\omega = x + ct$ the original local solution broke down to several smaller islands which freely propagate in time and space.

4.6 Periodic Noise

The last perturbation investigated is a periodic function $\eta = a\sin(\omega)$ and

$$-\nu f''(\omega) + f'(\omega)\left[c - \frac{\lambda}{2} f'(\omega)\right] - a\sin(\omega) = 0. \tag{16}$$

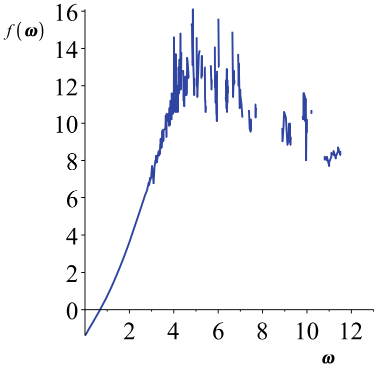


Fig. 11 The shape function for the Lorentzian noise. The applied parameters are $c_1 = 0.5, c_2 = 2, c = 1, \nu = 1, a = 1, \lambda = 3$

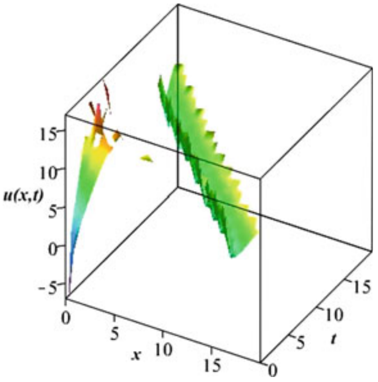


Fig. 12 The solution of the KPZ equation for Lorentzian noise, with the parameters mentioned above

The general solution can be given as

$$f(\omega) = \frac{1}{\lambda}\left(c\omega + 2\ln\left\{\frac{\lambda[c_1 C(\varepsilon_a) - c_2 S(\varepsilon_a)]}{\nu[-C'(\varepsilon_a)S(\varepsilon_a) + C(\varepsilon_a)S'(\varepsilon_a)]}\right\}\right), \tag{17}$$

where $C(\varepsilon_a)$, $S(\varepsilon_a)$, $C'(\varepsilon_a)$ and $S'(\varepsilon_a)$ are the Mathieu S and Mathieu C functions and the first derivatives. For basic properties we refer to [31]. For a complex study about Mathieu functions see [34–36]. In (17), we used the abbreviation of $\varepsilon_a = -\frac{c^2}{\nu^2}, -\frac{a\lambda}{\nu^2}, -\frac{\pi}{4} + \frac{\omega}{2}$.

Figure 13 shows a typical shape function for the periodic noise term. Due to the elaborate properties of even the single Mathieu C or S functions for some parameter pairs a, q the function is finite with periodic oscillations and for some neighboring parameters it is divergent for large arguments. No general parameter dependence can be stated. The parameter space of the set of six real values $(c_1, c_2, c, a, \nu, \lambda)$ is too large to map. After the evaluation of numerous shape functions we may state, that a typical shape function is presented with two larger islands close to the origin and numerous smaller intervals. For large argument ω the shape function shows a steep decay.

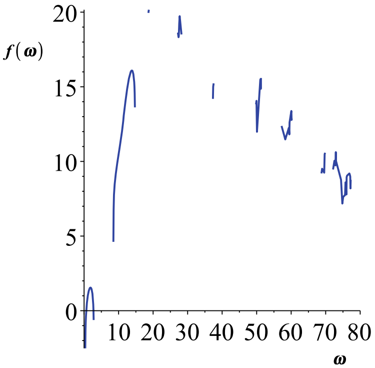


Fig. 13 The shape function for the periodic noise. The applied parameters are $c_1 = 0.5, c_2 = 2, c = 1, \nu = 1, a = 1, \lambda = 3$

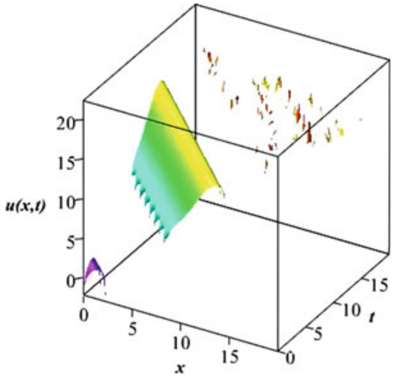


Fig. 14 The complete traveling wave solution $u(x, t)$ for periodic noise with the same parameter set as given above

Figure 14 shows the complete solution. Note, that the first two broader islands can be seen as they freely travel. Due to the finite resolution the smaller islands are represented as irregular noise in the background.

5 Conclusions

In summary, we can say that with an appropriate change of variables applying the traveling-wave Ansatz one may obtain analytic solution for the KPZ equation for one spatial dimension with numerous noise terms. We investigated four type of power-law noise ω^n with exponents of $-2, -1, 0, 1$, called the brown, pink, white and blue noise, respectively. Each integer exponent describes completely different dynamics. Additionally, the properties of Gaussian and Lorentzian noises are investigated. Providing completely dissimilar surfaces with growth dynamics. All solutions can be described with non-trivial combinations of various special functions, like error, Whittaker, Kummer or Heun. The parameter dependencies of the solutions are investigated and discussed. Future works are planned for the investigations of two dimensional surfaces.

Acknowledgements This work was supported by project no. 129257 implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme. The described study was carried out as part of the EFOP-3.6.1-16-00011 "Younger and Renewing University—Innovative Knowledge City—institutional development of the University of Miskolc aiming at intelligent specialization" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

References

1. Saito, Y.: Statistical Physics of Crystal Growth. World Scientific Press, Singapore (1996)
2. Kardar, M., Parisi, G., Zhang, Y.-C.: Phys. Rev. Lett. **56**, 889 (1986)
3. Barabási, A.-L.: Fractal Concepts in Surface Growth. Press Syndicate of the University of Cambridge, New York (1995)
4. Hwa, T., Frey, E.: Phys. Rev. A **44**, R7873 (1991)
5. Frey, E., Täubner, U.C., Hwa, T.: Phys. Rev. E **53**, 4424 (1996)
6. Lässig, M.: J. Phys.: Condens. Matter **10**, 9905 (1998)
7. Kriecherbauer, T., Krug, J.: J. Phys. A: Math. Theor. **43**, 403001 (2010)
8. Matsushita, M., Wakita, J., Itoh, H., Rafols, I., Matsuyama, T., Sakaguchi, H., Mimura, M.: Phys. A **249**, 517 (1998)
9. Kuramoto, Y., Tsuzki, T.: Prog. Theor. Phys. **55**, 356 (1976). Sivashinsky, G.I.: Phys. D, **4**, 227 (1982)
10. Kersner, R., Vicsek, M.: J. Phys. A: Math. Gen. **30**, 2457 (1997)
11. Sasamoto, T., Spohn, H.: Phys. Rev. Lett. **104**, 230602 (2010)
12. Calabrese, P., Doussal, P.L.: Phys. Rev. Lett. **106**, 250603 (2011)
13. Kelling, J., Ódor, G., Gemming, S.: Comput. Phys. Commun. **220**, 205 (2017)
14. Martynec, T., Klapp, S.H.L.: Phys. Rev. E **98**, 042801 (2018)
15. Sergi, D., Camarano, A., Molina, J.M., Ortona, A., Narciso, J.: Int. J. Mod. Phys. C **27**, 1650062 (2016)
16. Mello, B.A.: Phys. A **419**, 762 (2015)
17. Gilding, B.H., Kersner, R.: Progress in nonlinear differential equations and their applications. In: Travelling Waves in Nonlinear Diffusion-Convection Reactions. Birkhauser Verlag, Basel-Boston-Berlin (2004)
18. Cross, M.C., Hohenberg, P.C.: Rev. Mod. Phys. **65**, 851 (1993)
19. Van Saarloos, W.: Phys. Rep. **386**, 29 (2003)
20. He, J.H., Wu, X.H.: Chaos Solitons Fractals **30**, 700 (2006)
21. Aslan, I., Marinakis, V.: Commun. Theor. Phys. **56**, 397 (2011)
22. Benhamidouche, N.: Electron. J. Qual. Theory Diff. Equat. **15**, 1 (2008). http://www.math.u-szeged.hu/ejqtde/
23. Sedov, L.: Similarity and Dimensional Methods in Mechanics. CRC Press, Boca Raton (1993)
24. Barna, I.F.: Commun. Theor. Phys. **56**, 745 (2011)
25. Barna, I.F., László, M.: Chaos Solitons Fractals **78**, 249 (2015)
26. Campos, D.: Chapter 16. In: Handbook on Navier-Stokes Equations, Theory and Applied Analysis, pp. 275–304. Nova Publishers, New York, (2017)
27. Hopf, E.: Commun. Pure Appl. Math. **3**, 201 (1950)
28. Cole, J.D.: Quart. Appl. Math. **9**, 225 (1951)
29. Barna, I.F., Bognár, G., Guedda, M., Mátyás, L., Hriczó, K.: Math. Model. Anal. **25**(2) (2020)

30. Barna, I.F., Bognár, G., Guedda, M., Mátyás, L., Hriczó, K. https://arxiv.org/abs/1908.09615
31. Olver, F.W.J., Lozier, D.W., Boisvert, R.F., Clark, C.W.: NIST Handbook of Mathematical Functions. Cambridge University Press, Cambridge (2010)
32. Vallèe, O., Soares, M.: Airy Functions and Applications to Physics. World Scientific Publishing Company, Singapore (2004)
33. Ronveaux, A.: Heun's Differential Equations. Oxford University Press, Oxford (1995)
34. McLachlan, N.W.: Theory and Applications of Mathieu Functions. Dover, New York (1964)
35. Meixner, J., Schäfke, F.W.: Mathieusche Funktionen und Sphäroidfunktionen. Springer, Berlin (1954)
36. Ascott, F.M.: Periodic Differential Equations. Pergamon Press, Oxford (1964)

Triple Solutions for Elastic Beam Equations of the Fourth-Order with Boundary Conditions Subjected to an Elastic Device

G. Bonanno, A. Chinnì, and D. O'Regan

Abstract Under appropriate conditions on the nonlinear term, the existence of multiple solutions for a fourth-order problem is established. The result, obtained by variational techniques, is completed by applications and examples.

Keywords Fourth-order equations · Critical points · Variational methods

Mathematics Subject Classification 2010: 34B15

1 Introduction

In this paper, we will examine the following fourth-order problem

$$\begin{cases} u^{(iv)}(x) = \lambda f(x, u(x)) \ \text{ in } \ [0, 1] \\ u(0) = u'(0) = 0 \\ u''(1) = 0 \quad u'''(1) + \mu g(u(1)) = 0 \end{cases} \tag{$P_{\lambda,\mu}$}$$

G. Bonanno (✉) · A. Chinnì
Department of Engineering, University of Messina,
Contrada Di Dio, (S. Agata), 98166 Messina, Italy
e-mail: bonanno@unime.it

A. Chinnì
e-mail: achinni@unime.it

D. O'Regan
School of Mathematics, Statistics and Applied Mathematics,
National University of Ireland, Galway, Ireland
e-mail: donal.oregan@nuigalway.ie

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_20

where $f : [0, 1] \times \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are continuous functions and λ, μ are positive parameters. Solutions of problem $(P_{\lambda,\mu})$ are functions $u = u(x)$ which measure the downward deflection of a flexible elastic beam of length 1, clamped at its left end $x = 0$, resting on an elastic device (given by g) at its right end $x = 1$ and subjected to a deforming force f. Existence of solutions for problem $(P_{\lambda,\mu})$ has been extensively studied both with iterative methods and with variational methods. We refer the reader for example to [2–8].

The aim of this paper is to look for conditions in order that this problem admits multiple solutions. Precisely, using variational techniques, we will prove the existence of at least three classical solutions, by requiring a suitable behavior both of the nonlinear term and the elastic device. The main tool is a critical point theorem obtained by Bonanno and Marano in [1].

The paper consists of three sections. Section 2 contains some basic properties of the space in which we will search for solutions of problem $(P_{\lambda,\mu})$ and the technical features of the associated energy functional. The main result for problem $(P_{\lambda,\mu})$ and its proof are given in the same section. Section 3 is devoted to some applications and examples of the results illustrated in Sect. 2.

2 Existence of Three Solutions for Problem $(P_{\lambda,\mu})$

Denoted by $H^2([0, 1])$ the Sobolev space of all functions $u : [0, 1] \to \mathbb{R}$ such that u and its distributional derivative u' are absolutely continuous and u'' belongs to $L^2([0, 1])$, the solutions of problem $(P_{\lambda,\mu})$ lie in the following subspace

$$X := \{u \in H^2([0, 1]) : u(0) = u'(0) = 0\}\cdot$$

X is equipped with inner product

$$\langle u, v\rangle := \int_0^1 u''(t)v''(t)\, dt$$

and with norm

$$\|u\| := \left(\int_0^1 (u''(t))^2\, dt\right)^{\frac{1}{2}}$$

that make it an Hilbert space. As proved in [4], X is compactly embedded on $C^1([0, 1])$ and we have

$$\|u\|_{C^1([0,1])} := \max\left\{\|u\|_\infty, \|u'\|_\infty\right\} \leq \|u\| \tag{1}$$

for each $u \in X$. Fix $\lambda,\ \mu > 0$, and we consider the functional

$$I_{\lambda,\mu} := \Phi - \lambda \Psi_{\lambda,\mu},$$

where $\Phi, \Psi_{\lambda,\mu} : X \to \mathbb{R}$ are defined by

$$\Phi(u) := \frac{1}{2}\|u\|^2$$

and

$$\Psi_{\lambda,\mu}(u) := \int_0^1 F(x, u(x))\, dx + \frac{\mu}{\lambda} G(u(1))$$

for each $u \in X$ with $F(x,t) := \int_0^t f(x,\xi)\, d\xi$ and $G(t) := \int_0^t g(\xi)\, d\xi$ for each $x, t \in [0, 1]$. We list below some properties of the functionals Φ and $\Psi_{\lambda,\mu}$:

- Φ is sequentially weakly lower semicontinuous;
- Φ is coercive;
- Φ is in $C^1(X)$ and we have

$$\langle \Phi'(u), v \rangle = \int_0^1 u''(x)v''(x)\, dx$$

  for each $u, v \in X$;
- Φ' admits a continuous inverse on X^*
- $\Psi_{\lambda,\mu}$ is in $C^1(X)$ and we have

$$\langle \Psi'_{\lambda,\mu}(u), v \rangle = \int_0^1 f(x, u(x))v(x)\, dx + \frac{\mu}{\lambda} g(u(1))v(1)$$

  for each $u, v \in X$.

Weak solutions of problem ($P_{\lambda,\mu}$) coincide with critical points of the functional $I_{\lambda,\mu}$ for each $\lambda, \mu > 0$. In particular, in [4] the authors proved that, if $f : [0, 1] \times \mathbb{R} \to \mathbb{R}$ is continuous, then the critical points of $I_{\lambda,\mu}$ are classical solutions for problem ($P_{\lambda,\mu}$). The tool that will allow us to obtain multiple solutions for problem ($P_{\lambda,\mu}$) is the following critical point result obtained by G. Bonanno and S.A. Marano in [1]. In particular, for λ and μ in precise intervals, the existence of three classical solutions for problem ($P_{\lambda,\mu}$) is established.

Theorem 1. (Theorem 3.6 of [1]) *Let X be a reflexive real Banach space, $\Phi : X \to \mathbb{R}$ be a coercive, continuously Gâteaux differentiable and sequentially weakly lower semicontinuous functional whose Gâteaux derivative admits a continuous inverse on X^*, $\Psi : X \to \mathbb{R}$ be a continuously Gâteaux differentiable functional whose Gâteaux derivative is compact such that*

$$\inf_{x \in X} \Phi(x) = \Phi(0) = \Psi(0) = 0.$$

Assume that there exist $r > 0$ *and* $\bar{x} \in X$*, with* $r < \Phi(\bar{x})$*, such that:*

(a_1) $\dfrac{\sup\limits_{\Phi(x)\le r} \Psi(x)}{r} < \frac{\Psi(\bar{x})}{\Phi(\bar{x})}$;

(a_2) *for each* $\lambda \in \Lambda_r := \left]\dfrac{\Phi(\bar{x})}{\Psi(\bar{x})}, \dfrac{r}{\sup_{\Phi(x)\le r} \Psi(x)}\right[$ *the functional* $\Phi - \lambda\Psi$ *is coercive.*

Then, for each $\lambda \in \Lambda_r$*, the functional* $\Phi - \lambda\Psi$ *has at least three distinct critical points in* X*.*

Before introducing the main result, we make more precise some notation. With $\alpha \geq 0$, we put

$$F^\alpha := \int_0^1 \max_{|\xi|\le\alpha} F(x,\xi)\, dx$$

and

$$G^\alpha := \max_{|\xi|\le\alpha} G(\xi)\cdot$$

Theorem 2. *Assume that*

(f_1) *there exist two constants* $0 < \gamma < \delta$*, such that*

$$\frac{F^\gamma}{\gamma^2} < \frac{1}{8\pi^4}\left(\frac{3}{2}\right)^3 \frac{\int_{\frac{3}{4}}^1 F(x,\delta)\, dx}{\delta^2},$$

(f_2) $F(x,t) \geq 0$ *for almost every* $x \in [0,1]$ *and for all* $t \in [0,\delta]$*,*

(f_3) $\limsup\limits_{|t|\to+\infty} \dfrac{\sup_{x\in[0,1]} F(x,t)}{t^2} \leq 0.$

Then, for each $\lambda \in \Lambda_{\delta,\gamma} := \left]4\pi^4\left(\frac{2}{3}\right)^3 \dfrac{\delta^2}{\int_{\frac{3}{4}}^1 F(x,\delta)\, dx}, \dfrac{\gamma^2}{2F^\gamma}\right[$*, and for each* $g : \mathbb{R} \to \mathbb{R}$ *continuous such that* $l_g := \limsup_{|t|\to+\infty} \frac{G(t)}{t^2} < +\infty$*, there exists* $\tilde{\eta}_{\lambda,g} > 0$*, where*

$$\tilde{\eta}_{\lambda,g} = \begin{cases} \min\left\{\frac{\gamma^2 - 2\lambda F^\gamma}{2G^\gamma}, \frac{1}{2\max\{0,l_g\}}\right\} & \text{if } \; G(\delta) \geq 0 \\ \min\left\{\frac{\gamma^2 - 2\lambda F^\gamma}{2G^\gamma}, \frac{4\pi^4\delta^2 - \lambda\left(\frac{3}{2}\right)^3 \int_{\frac{3}{4}}^1 F(x,\delta)\, dx}{\left(\frac{3}{2}\right)^3 G(\delta)}, \frac{1}{\max\{0,2l_g\}}\right\} & \text{if } \; G(\delta) < 0, \end{cases} \tag{2}$$

such that for each $\mu \in]0, \tilde{\eta}_{\lambda,g}[$ *the problem* ($P_{\lambda,\mu}$) *admits at least three classical solutions.*

Proof. Fix $\lambda \in \Lambda_{\delta,\gamma}$. Taking into account that $G^{\gamma} := \max_{|\xi|\leq\gamma} G(\xi) \geq G(0) = 0$, we observe that $\tilde{\eta}_{\lambda,g} > 0$. Indeed, if $G(\delta) \geq 0$, by $\lambda \in \Lambda_{\delta,\gamma}$ it follows that $\gamma^2 - 2\lambda F^{\gamma} > 0$. Hence $\tilde{\eta}_{\lambda,g} > 0$. Let $G(\delta) < 0$. We have by $\lambda \in \Lambda_{\delta,\gamma}$ that

$$4\pi^4 \left(\frac{2}{3}\right)^3 \frac{\delta^2}{\int\limits_{3/4}^{1} F(x,\delta)dx} < \lambda,$$

which implies $4\pi^4\delta^2 - \lambda \left(\frac{3}{2}\right)^3 \int\limits_{3/4}^{1} F(x,\delta)dx < 0$. Hence $\tilde{\eta}_{\lambda,g} > 0$, in this case as well. In both cases we have $\tilde{\eta}_{\lambda,g} = +\infty$ when $G^{\gamma} = 0$ and/or $l_g \leq 0$.

Now, fix $g : \mathbb{R} \to \mathbb{R}$ continuous and $\mu \in]0, \tilde{\eta}_{\lambda,g}[$, and we apply Theorem 1 to the functionals Φ,$\Psi_{\lambda,\mu}$ defined on X. As stated before, Φ and $\Psi_{\lambda,\mu}$ verify the regularities requested in Theorem 1. We prove that assumptions of Theorem 1 are verified for $r = \frac{\gamma^2}{2}$ and $\bar{v} \in X$ defined by

$$\bar{v}(x) = \begin{cases} 0 & x \in [0, \frac{3}{8}] \\ \delta \cos^2(\frac{4\pi x}{3}) & x \in]\frac{3}{8}, \frac{3}{4}[\\ \delta & x \in [\frac{3}{4}, 1]. \end{cases} \tag{3}$$

Indeed we have

$$\Phi(\bar{v}) = 4\pi^4\delta^2 \left(\frac{2}{3}\right)^3. \tag{4}$$

On the other hand one has

$$\Psi_{\lambda,\mu}(\bar{v}) = \int_0^1 F(x, \bar{v}(x))\, dx + \frac{\mu}{\lambda} G(\delta) \geq \int_{\frac{3}{4}}^1 F(x,\delta)\, dx + \frac{\mu}{\lambda} G(\delta)$$

where we applied condition (f_2) to ensure $\int_0^1 F(x, \bar{v}(x))\, dx \geq \int_{\frac{3}{4}}^1 F(x,\delta)\, dx$.
So, we obtain

$$\frac{\Psi_{\lambda,\mu}(\bar{v})}{\Phi(\bar{v})} \geq \frac{\int_{\frac{3}{4}}^1 F(x,\delta)\, dx + \frac{\mu}{\lambda} G(\delta)}{4\pi^4\delta^2 \left(\frac{2}{3}\right)^3}. \tag{5}$$

Condition (1) provides that $\|u\| \leq \gamma = \sqrt{2r}$ and $\|u\|_{\infty} \leq \gamma$ for each $u \in \Phi^{-1}(]-\infty, r])$ and so we have

$$\frac{1}{r} \sup_{u \in \Phi^{-1}(]-\infty,r])} \Psi_{\lambda,\mu}(u) \leq \frac{2}{\gamma^2} F^{\gamma} + \frac{2}{\gamma^2} \frac{\mu}{\lambda} G^{\gamma}. \tag{6}$$

Now, if $G(\delta) \geq 0$, taking into account that $\lambda \in \Lambda_{\delta,\gamma}$ and that in particular $\tilde{\eta}_{\lambda,g} \leq \frac{\gamma^2 - 2\lambda F^\gamma}{2G^\gamma}$, one has

$$\frac{2}{\gamma^2}F^\gamma + \frac{2}{\gamma^2}\frac{\mu}{\lambda}G^\gamma \leq \frac{2}{\gamma^2}F^\gamma + \frac{2}{\gamma^2}\frac{\tilde{\eta}_{\lambda,g}}{\lambda}G^\gamma \leq \frac{1}{\lambda}$$

and

$$\frac{1}{\lambda} < \frac{1}{4\pi^4\delta^2}\left(\frac{3}{2}\right)^3 \int_{\frac{3}{4}}^1 F(x,\delta)\,dx \leq \frac{1}{4\pi^4\delta^2}\left(\frac{3}{2}\right)^3 \left(\int_{\frac{3}{4}}^1 F(x,\delta)\,dx + \frac{\mu}{\lambda}G(\delta)\right).$$

If $G(\delta) < 0$, taking into account that, in particular,

$$\tilde{\eta}_{\lambda,g} \leq \min\left\{\frac{\gamma^2 - 2\lambda F^\gamma}{2G^\gamma}, \frac{4\pi^4\delta^2 - \lambda\left(\frac{3}{2}\right)^3 \int_{\frac{3}{4}}^1 F(x,\delta)\,dx}{\left(\frac{3}{2}\right)^3 G(\delta)}\right\}, \tag{7}$$

we have as above

$$\frac{2}{\gamma^2}F^\gamma + \frac{2}{\gamma^2}\frac{\mu}{\lambda}G^\gamma \leq \frac{2}{\gamma^2}F^\gamma + \frac{2}{\gamma^2}\frac{\tilde{\eta}_{\lambda,g}}{\lambda}G^\gamma \leq \frac{1}{\lambda}$$

and, again from (7),

$$\frac{1}{\lambda} < \frac{1}{4\pi^4\delta^2}\left(\frac{3}{2}\right)^3 \left(\int_{\frac{3}{4}}^1 F(x,\delta)\,dx + \frac{\mu}{\lambda}G(\delta)\right).$$

In all cases, taking into account (5) and (6), we obtain

$$\frac{1}{r}\sup_{u\in\Phi^{-1}(]-\infty,r])} \Psi_{\lambda,\mu}(u) < \frac{1}{\lambda} < \frac{\Psi_{\lambda,\mu}(\bar{v})}{\Phi(\bar{v})},$$

and so assumption (a_1) of Theorem 1 is verified.
Moreover, we observe that from $0 < \gamma < \delta$, one has $\Phi(\bar{v}) > r$. Indeed, if $4\pi^4\delta^2 \left(\frac{2}{3}\right)^3 = \Phi(\bar{v}) \leq r = \frac{\gamma^2}{2}$ then $\delta < \frac{3\sqrt{3}}{8\pi^2}\gamma < \gamma$, a contradiction with the hypothesis.
Now, we prove that the functional $I_{\lambda,\mu}$ is coercive.
Fix $\max\{0, l_g\} < l < \frac{1}{2\mu}$, and there exists a positive constant k such that

$$G(\xi) \leq l\xi^2 + k$$

for each $\xi \in \mathbb{R}$. Now, fix $0 < \varepsilon < \dfrac{\frac{1}{2} - \mu l}{\lambda}$. From (f_3) there is a positive constant k_ε such that

$$F(x,\xi) \leq \varepsilon\xi^2 + k_\varepsilon$$

for each $(x, \xi) \in [0, 1] \times \mathbb{R}$. Taking into account that $\|u\|_\infty \leq \|u\|$, it follows that, for each $u \in X$,

$$I_{\lambda,\mu}(u) = \Phi(u) - \lambda\Psi_{\lambda,\mu}(u) \geq (\frac{1}{2} - \lambda\varepsilon - l\mu)\|u\|^2 - \lambda k_\varepsilon - \mu k.$$

This leads to the coercivity of $I_{\lambda,\mu}$ and condition (a_2) of Theorem 1 is also verified. Finally, since $\lambda \in \Lambda_{\delta,\gamma} \subseteq \left]\frac{\Phi(\bar{v})}{\Psi_{\lambda,\mu}(\bar{v})}, \frac{r}{\sup_{\Phi(u)\leq r} \Psi_{\lambda,\mu}(u)}\right[$, Theorem 1 guarantees the existence of three distinct critical points for the functional $I_{\lambda,\mu}$ that are classical solutions for problem $(P_{\lambda,\mu})$. □

3 Applications and Examples

The following result is a consequence of Theorem 2 and concerns the autonomous case.

Corollary 1. *Assume that* $f : \mathbb{R} \to [0, +\infty[$, $f \not\equiv 0$ *is a continuous function such that*

$(\tilde{f}_1)$ $\lim_{t\to 0^+} \frac{f(t)}{t} = \lim_{t\to +\infty} \frac{f(t)}{t} = 0.$

Then, for each $\lambda > \bar{\lambda} := 16\pi^4(\frac{2}{3})^3 \inf\{\delta > 0 : \frac{\delta^2}{F(\delta)} > 0\}$, *for each* $g : \mathbb{R} \to [0, +\infty[$ *continuous such that* $\lim_{t\to 0^+} \frac{g(t)}{t} = \lim_{t\to +\infty} \frac{g(t)}{t} = 0$ *and for each* $\mu > 0$, *the problem*

$$\begin{cases} u^{(iv)}(x) = \lambda f(u(x)) \text{ in } [0, 1] \\ u(0) = u'(0) = 0 \\ u''(1) = 0 \quad u'''(1) + \mu g(u(1)) = 0 \end{cases} \tag{8}$$

admits at least three classical solutions.

Proof. First of all we observe that nonnegativity of f implies that F is not decreasing on $\mathbb{R}$ and in particular that $F(t) \geq 0$ for $t \in [0, +\infty[$, so condition (f_2) requested in Theorem 2 is clearly verified. Moreover, condition (f_3) of Theorem 2 follows clearly by $(\tilde{f}_1)$.

Fix $\lambda > \bar{\lambda}$, $g : \mathbb{R} \to [0, +\infty[$ continuous such that $\lim_{t\to 0^+} \frac{g(t)}{t} = \lim_{t\to +\infty} \frac{g(t)}{t} = 0$ and $\mu > 0$, and taking into account that $F^\alpha = F(\alpha)$ and $G^\alpha = G(\alpha)$ for each $\alpha \geq 0$, we choose $\bar{\delta} > 0$ such that

$$\frac{1}{\lambda} < \frac{F(\bar{\delta})}{\bar{\delta}^2} \left(\frac{3}{2}\right)^3 \frac{1}{16\pi^4}.$$

Because we have

$$\lim_{t\to 0^+} \frac{t^2 - 2\lambda F(t)}{2G(t)} = +\infty,$$

one has in particular $\frac{t^2-2\lambda F(t)}{2G(t)} > \mu$ for $t \in]0, \gamma_1[$, while $(\tilde{f_1})$ ensures

$$\frac{F(t)}{t^2} < \frac{1}{2\lambda}$$

for $t \in]0, \gamma_2[$. With $\bar{\gamma} < \min\{\bar{\delta}, \gamma_1, \gamma_2\}$, we have $\bar{\gamma} < \bar{\delta}$,

$$\frac{F(\bar{\gamma})}{\bar{\gamma}^2} < \frac{1}{32\pi^4}\left(\frac{3}{2}\right)^3 \frac{F(\bar{\delta})}{\bar{\delta}^2}$$

and, taking into account that $l_g = 0$ and $G(\bar{\delta}) \geq 0$,

$$0 < \mu < \frac{\bar{\gamma}^2 - 2\lambda F(\bar{\gamma})}{2G(\bar{\gamma})} = \tilde{\eta}_{\lambda,g}.$$

So, all the conditions requested in Theorem 2 are verified and problem (8) admits at least three classical solutions. □

Remark 1. In Corollary 1, since we have $f(0) = g(0) = 0$ then problem (8) admits at least two non trivial classical solutions.

Example 1. The function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(t) := \begin{cases} |t|^3, & |t| \leq 1, \\ \sqrt{|t|}, & |t| > 1 \end{cases} \tag{9}$$

verifies for example the assumptions of Theorem 1.

Acknowledgements This study was partly funded by Research project of MIUR (Italian Ministry of Education, University and Research) Prin 2017 “Nonlinear Differential Problems via Variational, Topological and Set-valued Methods” (Grant Number: 2017AYM8XW). The first two authors are members of the Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. Bonanno, G., Marano, S.A.: On the structure of the critical set of non-differentiable functions with a weak compactness condition. Appl. Anal. **89**, 1–10 (2010)
2. Bonanno, G., Chinnì, A., Tersian, S.: Existence results for a two point boundary value problem involving a fourth-order equation. Electron. J. Qual. Theor Diff. Equ. **33**, 1–9 (2015)
3. Galewski, M.: On the Dirichlet problem for a nonlinear elastic beam equation. Appl. Math. Comput. **217**, 4295–4305 (2010)
4. Yang, L., Chen, H., Yang, X.: The multiplicity of solutions for fourth-order equations generated from a boundary condition. Appl. Math. Letters **24**, 1599–1603 (2011)
5. Cabada, A., Tersian, S.: Multiplicity of solutions of a two point boundary value problem for a fourth-order equation. Appl. Math. Comput. **24**, 1599–1603 (2011)

6. Grossinho, M.R., Tersian, S.: The dual variational principle and equilibria for a beam resting on a discontinuous nonlinear elastic foundation. Nonlinear Anal. **41**, 417–431 (2000)
7. Ma, T.F., da Silva, J.: Iterative solutions for a beam equation with nonlinear boundary conditions of third order. Appl. Math. Comput. **159**, 11–18 (2004)
8. Ma, T.F.: Positive solutions for a beam equation on a nonlinear elastic foundation. Math. Comput. Model. **39**, 1195–1201 (2004)

Dynamics of Discrete Operator Equations

George L. Karakostas

Abstract The translation of nonautonomous difference causal operators defined on the set of sequences is introduced and some facts from the dynamics of abstract nonautonomous difference equations are presented. Also, a sufficient number of applications are given to obtain results on asymptotic stability of the equilibria.

1 Introduction

Autonomous (continuous or discrete) systems are invariant in time. Any action in nonautonomous system depends on time. We are interesting in discrete systems. Such systems arise, mainly, by taking discrete approximation of continuous models and they are described by the so called difference equations. They are used to obtain approximate solutions of mathematical problems with recurrences, or to build various discrete models in economics, psychology, sociology, etc. For instance, the continuous time model for the well known logistic equation $dN/dt = rN(1 - N/K)$, has discrete analogs the equations $N_{n+1} = N_n + RN_n(1 - N_n/K)$, $N_{n+1} = N_n \exp(R(1 - N_n/K))$ and others. Nevertheless, such equations may arise independently, such as $x_{n+1} = \beta + 1/x_n x_{n-1}$ [8]. Perhaps the first theoretical results on nonautonomous difference equations were given by Poincaré [1885] and Peron [1921], see the references in [16], pp. 343–344. The Poincare and Perron Theorems provide information for the asymptotic behavior of the quantities x_{n+1}/x_n, or $(x_n)^{1/n}$, when (x_n) is a solution of a nonautonomous linear difference equation. Goldberg's book [25] (1986)[1] is a good source of examples on difference equations. Also, the book by LaSalle, [42] published in 1976, deals, mainly, with the role of Lyapunov functions to the study of stability in autonomous systems, while in [43] a more general situation is investigated. A deep exhibition of ergodic theory related to nonautonomous dynamical systems can be found in Petersen's book [47]. Elaydi's book [16] is an extended presentation of autonomous and nonautonomous difference equations. A very interesting investigation with a great number of examples and a rich bibliography of difference equations is exhibited in the Agarwal's book [6], see, also, [4].

G. L. Karakostas (✉)
Department of Mathematics, University of Ioannina, 451 10 Ioannina, Greece
e-mail: gkarako@uoi.gr; gkarako@hotmail.com

[1] The first edition of the book goes back to 1958.

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_21

Most of the known works dealing with this topic refer to autonomous difference equations, due to their invariance in time and the fact that the solutions can be formally explicitly given by the composition of the response function. Indeed, for the difference equation $x_{n+1} = f(x_n)$ the solution is, trivially, given by $x_n = f^{(n)}(x_0)$, where x_0 is the initial state. So, what one has to do is to obtain the iteration of f. Notice that in this case one can predict chaos, see the Sarkovskii's Theorem [21], or the Li-Yorke relation of periodicity with chaos [44], as well as the books by Abraham, Gardini and Mira [1], Bahi and Guyeux [9], Zhang [54], Devaney [12], etc. On the other hand the nonautonomous case is more complicated. Indeed, let us borrow from [24] (pp. 49–50) the simple nonhomogeneous linear difference equation

$$x_{n+1} = a_n x_n + b_n. \tag{1}$$

Although the solution can be given in a closed form as

$$x_n = x_0 \prod_{k=0}^{n-1} a_k + \sum_{k=0}^{n-2} b_k \prod_{i=k+1}^{n-1} a_i + b_{n-1}$$

in general, no asymptotic properties of it can easily be found. In Sect. 5 we shall return to it.

Generally, if we have a sequence of mappings $Q^n : X \to X$, where X is a metric space, then the difference equation (or process)

$$x_{n+1} = Q^n(x_n) \tag{2}$$

admits the solution $x_n = Q^{n-1} \circ Q^{n-2} \circ \cdots \circ Q^0(x_0)$, where x_0 is the initial value. However the problem of finding the composition of all these operators and investigating the chain transitivity and attractivity by critical values is not so easy. See, e.g., [53]. The problem becomes more difficult if the dependence of the operator Q^n is not only on x_n, but on all, or some previous values of the sequence (x_n). For instance, the Volterra difference equation $x_{n+1} = \sum_{j=0}^{n} K(n, j, x_j)$, studied in the literature (see, e.g. [19] and the references therein), or equations described in Sect. 2 of [5], cannot be written in the form (2). However though the problem gets a simple form, its complexity stays on. Therefore, in order to see the asymptotic behavior of the solutions we have to apply other methods based on the sense of dynamical systems.[2] To this approach in a series of papers Elaydi and Sacker (see [18] and the references therein) presented a skew-product semi-flow, as the discretization of the continuous dynamical systems, as they were suggested by Sell [49]. (The meaning of the skew-product flow appeared in a paper of Miller [45] in 1965.) In this work we extend their main idea. In some of our previous works [35, 36] we have studied the asymptotic behavior of autonomous difference equations by using the so

[2] Linearization is a good method, but, notice that sometimes it gives false information, see, [20], p. 22.

called full limiting sequences method, suggested in 1989, [29]. From the first point of view the method seemed to be very simple, but latter it was proved to be very useful (see, e.g., [30–34, 37, 41]. We shall recall the notion of the limiting and full limiting sequences in subsequent sections. What is new and it is introduced here is the meaning of the translation of causal operators defined on a set of sequences. Then a skew-product semi-flow is build by using the shifting process of sequences and the translation of the operator. These notions are used to obtain the full limiting equations of difference equations and predict results on the asymptotic behavior of several specific types of nonautonomous difference equations. The work is organized as follows: Section 2 contains a background on the shifting semi-flow of sequences with some facts on limiting and full limiting equations. Section 3 presents the translation and the dynamics of nonautonomous and discrete operators defined on the set of d-dimensional sequences. The full limiting orbits and limiting equations are exhibited in Sect. 4, while in Sect. 5 and 6 several applications to some discrete models are given.

2 Full Limiting Sequences

We begin with some facts which we borrow from our previous research (see, [29]) and refer to the so called shifting semi-flow, present the theory of full limiting sequences and give a short review of some of its applications. Our approach is quite different from the usual arguments on point-shifting discrete dynamical systems exhibited elsewhere in the literature, see, e.g. Easton [14], LaSalle [42], etc.

Let $\mathbf{Z}$ be the set of all integers, $\mathbf{N}$ the set of positive integers and $\mathbf{N}_0$ the set $\mathbf{N} \cup \{0\}$. We shall work on a metric space (X, d). In some cases we need X to be linear. Let $E := X^{\mathbf{N}_0}$. We assume that E is endowed with the point-wise convergence, or equivalently, convergence uniformly on finite sets.

Let $\mathscr{R}(x)$ be the range of a sequence $x \in E$.

For any $n \in \mathbf{N}_0$ and $x \in E$ define the *s*hifting operator $S : E \to E$ by the type $(Sx)_k = x_{k+1}, \quad k = 0, 1, 2, \cdots$. Then for each n we set $S^n x = S(S^{n-1}x)$, where S^0 is the identity operator. Thus, we have $(S^n x)_k = x_{k+n}$. Later on, in case we have a two sided sequence x we shall permit the index n to be negative. Clearly, S^n maps E into itself, it is continuous and furthermore the mapping $\pi : (n, x) \to S^n x$ defines a discrete semi-dynamical system with phase-space E, called *t*he shifting semi-flow.

The proof of the next result is obvious.

Proposition 1. *A trajectory $\pi(\cdot, x)$ is (i) stationary (m-periodic), if and only if the sequence x is constant (m periodic), (ii) compact (or Lagrange stable) with respect to the semi-flow π if and only if the set $\mathscr{R}(x)$ is relatively compact.*

Proposition 2. *Take $x \in E$ with $\mathscr{R}(x)$ relatively compact. Then the ω-limit set $\omega(x)$ with respect to the semi-flow π is nonempty, compact, invariant and invariantly connected.*[3] *Also any $\bar{x} \in \omega(x)$ has a full trajectory $U(\bar{x})$ which stays in $\omega(x)$.*

Proof. From Proposition 1, the trajectory $\pi(\cdot, x)$ is compact. Then the first four facts follow easily as in the general theory on discrete dynamical systems (see e.g. [42]). For the last one take any $\bar{x} \in \omega(x)$. Then there is a sequence of positive integers (n_m) such that $\lim \pi(n_m, x) = \bar{x}$. Since $\mathscr{R}(x)$ is relatively compact, there is a subsequence $n_{m,1}$ of (n_m), such that the limit $\lim_m x_{-1+n_{m,1}} =: l_1$ exists. Similarly, there is a subsequence n_{m_2} of $(n_{m,1})$, such that the limit $\lim_m x_{-2+n_{m,2}} =: l_2$ exists. We proceed inductively for any positive integer k to obtain a subsequence $n_{m,k}$ of $(n_{m,k-1})$, such that the limit $\lim x_{-k+n_{m,k}} =: l_k$ exists. Finally define $\hat{x}_{-k} =: l_k$ and $\hat{x}_k =: \bar{x}_k$ for each nonnegative integer k. Obviously, the so defined item $\hat{x}$ is a two-sided sequence $\cdots, \hat{x}_{-k}, \hat{x}_{-k+1}, \cdots, \hat{x}_0, \hat{x}_1, \hat{x}_2, \cdots$ whose the restriction on $\mathbf{N}_0$ is the sequence $\bar{x}$. The full trajectory $U(\bar{x})$ of $\bar{x}$ is defined by $U(\bar{x})(k) := S^k\bar{x}$ and it is the set $\{u \in E : (\exists k \in \mathbf{Z})\ u_n = \hat{x}_{k+n},\ \ (\forall n \in \mathbf{N}_0)\}$. It remains to show that any term of $U(\bar{x})$ stays in $\omega(x)$. To this end, take any $u \in U(\bar{x})$. Then we have $U(\bar{x})(k)_n = u_n = \hat{x}_{k+n}$, for a certain $k \in \mathbf{Z}$. If $k \geq 0$, then we have $u = \lim_m \pi(k + n_m, x)$. If $k < 0$, then for any $s > -k$, we have $n_{m,s} \geq m_s \geq s > -k$ and so $u = \lim_m \pi(k + n_{m,s}, x)$, which proves the result.

We call $\bar{x}$ a *limiting sequence* of x and $\hat{x}$ a *full limiting sequence* of x. Since the limit sets are invariant, it follows that any limiting or full limiting sequence of a limiting or full limiting sequence of a sequence (x_n) is, again, a limiting, or full limiting, respectively, sequence of (x_n).

If a sequence converges to a limit, say $l \in X$, then the set of full limiting sequences consists of the two-sided constant sequence $\cdots, l, l, \cdots$.

A sequence $x =: (x_n)$, is called *slowly varying*, if for each positive integer k it holds $\lim_{n\to+\infty} d(x_{n+k}, x_n) = 0$. Thus, if it is bounded then, any of its full limiting sequences is constant.

We remind that a sequence (x_n) is *almost periodic*, if for each $\epsilon > 0$ the set of all $p \in \mathbf{Z}$ for which $d(x_{n+p}, x_n) \leq \epsilon$, for all $n = 0, 1, 2, \cdots$ is relatively dense, in the sense that there is a m such that every interval of positive integers of length m contains at least one such p. (Such a number p is called ϵ-period.) Some interesting facts about the construction of almost periodic sequences as well as examples of almost periodic sequences can be found in the literature, see, e.g. [51, 52] and the references therein.

Motivated from the notion of an asymptotically almost periodic function (introduced by Fréchet in 1941), a bounded sequence (x_n) is called *asymptotically (Bohr-) almost periodic*, if for every $\epsilon > 0$ we can find integers $l > 0$ and $M > 0$ such that every subinterval J of positive integers of length l contains at least one number m such that $d(x_{n+m}, x_n) \leq \epsilon$, for all $n \geq M$. According to [10], Theorem 1.3.1, if a point of a metric space is asymptotically almost periodic with respect to a semi-flow,

[3] This means that the set cannot be written as the disjoint union of two closed invariant sets.

then its ω limit set coincides with the closure of the almost periodic trajectory. In the spirit of this fact we state the following obvious result:

Proposition 3. *If a sequence (x_n) is asymptotically almost periodic, then any full limiting sequence is almost periodic.*

For the inverse of the result and more facts about the asymptotic almost periodicity and Poisson stability will be given in a forthcoming work.

The main applications of this theory is based on the following fact: If x is a bounded sequence then any sequence of positive integers converging to $+\infty$ produces a full limiting sequence. To apply this fact consider an autonomous delay difference equation of the form

$$x'_n := x_{n+1} = H(x_n, x_{n-1}, \cdots, x_{n-k}). \tag{3}$$

where $H : (X)^{k+1} \to X$.

It is easy to see that if $x \in E$ is a solution of (3), then $S^m x$ is also a solution for any $m \geq 0$. Moreover, if $\hat{x}$ is a full limiting sequence of x, we have $\hat{x}_n = \lim_l \pi(n, \pi(r_l, x))_0 = \lim_l \pi(r_l + n, x)_0 = \lim_l x_{r_l+n}$, for each $n \in \mathbf{Z}$, and some sequence (r_l) converging to $+\infty$. (Clearly $r_l + n \geq 0$, for all large l.) Thus, from (3) we see that $x'_{r_l+n} = H(x_{r_l+n}, x_{r_l+n-1}, \cdots, x_{r_l+n-k})$ and, passing to the limits, we get $\hat{x}'_n = H(\hat{x}_n, \hat{x}_{n-1}, \cdots, \hat{x}_{n-k})$. Therefore, any full limiting sequence of a solution of (3) satisfies the same equation for all integers. Such a sequence (if it exists) is usually called *a full solution*.

A great number of applications of the theory of full limiting sequences are presented in the literature. See, e.g. [23, 26, 29–37, 46].

3 Dynamics of Nonautonomous Discrete Operators

Let x, u be two sequences in E. For any $m \in \mathbf{N}_0$ define the new sequence $\left(\mu_{m,x} u\right)_k := x_k,\ \ k < m$ and $\left(\mu_{m,x} u\right)_k := u_{k-m},\ \ k \geq m$. Let D be an open subset of E with the following property:

For any x, u in D and $m \in \mathbf{N}_0$ the sequence $\mu_{m,x} u$ belongs to D.

Let $\mathscr{D}$ be the class of all such domains. We shall work on the set $\mathscr{T}$ of all operators with domain in $\mathscr{D}$ which are causal in the sense that, for any m, if it holds $x_n = y_n$, for all of $n = 0, 1, 2, \cdots, m$, then $(Tx)_m = (Ty)_m$. The class $\mathscr{T}$ is endowed with the following continuous convergence structure (which is inspired from Artstein's Appendix of [42] as well as from the more general convergence structure applying in [28]: A sequence of operators $T_k \in \mathscr{T}$ with domains D_{T_k} $k = 0, 1, 2, \cdots$, will converge to an operator $T \in \mathscr{T}$, with domain D_T, if the following facts hold: *For any sequence (u^k) such that $u^k \in D_{T_k}, k = 1, 2, \ldots$ and $u^k \to u$, for a certain $u \in E$, it holds $u \in D_T$ and $T_k u^k \to Tu$.*

The prototype of such a causal operator is the discrete Volterra operator, i.e., the equation $x_{i+1} = F(i, x_0, \cdots, x_i),\ \ i \geq 0$, discussed elsewhere (see, e.g., [11]), but we will restrict ourselves to the familiar discrete version of the Volterra integral equation[4] defined by

$$(Tx)_m = a_m + \sum_{j=0}^{m} g_{m,j}(x_j), \;\; m = 0, 1, 2, \cdots . \tag{4}$$

If U is the common domain of the family $(g_{m,j})$, then D_T is the set of all $x \in E$ such that $x_k \in U$, for all $k = 0, 1, 2, \ldots$.

Assume that the metric space X is linear and let $T \in \mathscr{T}$. For any $x \in D_T$ and $m \in \mathbf{N}_0$ define the *translation of the operator T along the sequence x by m* by the type

$$(T_{m,x}u)_k = (Sx)_m - (Tx)_m + (T\mu_{m,x}u)_{m+k}, \;\; k = 0, 1, \cdots , \;\; u \in D_T.$$

If x is a solution of the operator equation $Sx = Tx$, or X is a not necessarily linear metric space, then *the translation of T along a solution x at any m* is defined by

$$(T_{m,x}u)_k = (T\mu_{m,x}u)_{m+k}, \;\; k = 0, 1, \cdots , \;\; u \in D_T. \tag{5}$$

In this case we have $T_{0,x} = T$. Therefore in the linear case, the factor $(Sx)_m - (Tx)_m$ denotes the perturbation of the translation when x is not a solution. For example, the translation of the operator T defined by (4) along a solution x of equation $Sx = Tx$, at any m, is given by $(T_{m,x}u)_k = a_{m+k} + \sum_{j=-m}^{-1} g_{m+k,m+j}((Sx)_{m+j}) + \sum_{j=0}^{k} g_{m+k,m+j}(u_j)$.

Theorem 1. *For any $T \in \mathscr{T}$, $m \in \mathbf{N}_0$ and $x \in \mathscr{D}_T$ the translation operator $T_{m,x}$ has the following properties: (i) $T_{m,x} \in \mathscr{T}$. (ii) The operators $T_{m+l,x}$ and $(T_{m,x})_{l,S^m x}$ have the same domain and are identically equal. (iii) For each m the mapping $x \to T_{m,x}$ is continuous.*

Proof. (i) This is implied from the definition of the translation.

(ii) It is easy to see that for all u with $u_0 = x_{m+l}$ it holds $(\mu_{m,x}\mu_{l,S^m x}u)_s = (\mu_{m+l,x}u)_s$. Therefore $((T_{m,x})_{l,S^m x}u)_k = (S^{m+1}x)_l - (T_{m,x}S^m x)_l + (T_{m,x}\mu_{l,S^m x} u)_{l+k} = (Sx)_{l+m} - [(Sx)_m - (Tx)_m + (Tx)_{l+m}] + (Sx)_m - (Tx)_m + (T\mu_{m,x}\mu_{l,S^m x} u)_{m+l+k} = (Sx)_{l+m} - (Tx)_{l+m} + (T\mu_{m+l,x}u)_k = (T_{m+l,x}u)_k$. Property (iii) is a consequence of the continuity of the operator T.

Assume that X is a linear space. Given $T \in \mathscr{T}$ and $x \in D_T$ define the mapping $\gamma : (m; (x, T_{0,x})) \to (\pi(m, x), T_{m,x})$, with domain $\mathscr{W} := \mathbf{N}_0 \times \{(x, S) : \;(\exists T \in \mathscr{T}) : x \in D_T, S = T_{0,x}\}$. If X is not necessarily linear space, define the mapping γ, as above, with domain $\mathscr{W} := \mathbf{N}_0 \times \{(x, T) : \; T \in \mathscr{T}, \; x \in D_T, \; Sx = Tx\}$. It is easy

[4] The background for discrete Volterra equations can be found in the well-known monograph by Agarwal [3], as well as in Elaydi [17] and Kocić and Ladas [40].

to see that γ is a (discrete) skew-product semi-flow.[5] Indeed, continuity is obvious and the identity condition $\gamma(0; (x, T_{0,x})) = (\pi(0, x), T_{0,x}) = (x, T_{0,x})$ as well as the cocycle condition $\gamma(m; \gamma(l; (x, T_{0,x})) = \gamma(m; (\pi(l, x), T_{l,x})) = (\pi(m; \pi(l; x)),$ $(T_{m,x})_{m,\pi(l;x)} = (\pi(m + l; x), T_{m+l,x})$ hold because of Theorem 1. Continuity of $(x, T) \to \gamma(m; (x, T))$ is obvious.

Theorem 2. *Assume that X is a linear metric space, $T \in \mathscr{T}$ and $x \in D_T$. A point $(x, T_{0,x})$ is m-periodic with respect to the semi-flow γ, if and only if it holds*

$$S^m x = x \ \text{ and } \ TS^m u = S^m Tu + (Tx)_0 - (Tx)_m, \tag{6}$$

for all $u \in D_T$ with $u_k = x_k$, $k = 0, 1, \cdots, m$.

Proof. If (x, T) is periodic with period m, it holds $S^m x = x$ and $T_{m,x} = T_{0,x}$. From the first we get $x_{m+k} = x_k$ for all $k = 0, 1, 2, \cdots$. Let $u \in E$ with $u_k = x_k$, $k = 0, 1, \cdots, m$. Then we have $(T_{m,x} S^m u)_k = (T_{0,x} S^m u)_k$, $k = 0, 1, 2, \cdots$. The first part is equal to $(Sx)_m - (Tx)_m + (T\mu_{m,x} S^m u)_{m+k} = (Sx)_m - (Tx)_m + (Tu)_{m+k} = x_{m+1} - (Tx)_m + (S^m(Tu))_k$, while the right part is equal to $(Sx)_0 - (Tx)_0 + (T\mu_{0,x} S^m u)_k = x_1 - (Tx)_0 + (TS^m u)_k$. This shows the "only if" part.

To show the "if" part, fix any $u \in E(x_0)$, $k \in \mathbf{N}_0$ and define $y := \mu_{m,x}[\mu_{k,S^m x} u]$. From (6) we have $S^m x = x$, thus $y = \mu_{m,x}[\mu_{k,x} u]$. It is clear that, for any $s \leq m$, it holds $y_s = x_s$. Hence from (6) we get

$$(TS^m y)_{s+k} = (S^m Ty)_{s+k} + (Tx)_0 - (Tx)_m, \ \ s = 0, 1, \cdots. \tag{7}$$

On the other hand relation (6) implies that

$$(Tx)_0 - (Tx)_m = (TS^m x)_k - (S^m(Tx))_k = (Tx)_k - (Tx)_{m+k}. \tag{8}$$

Now we observe that for any $s = 0, 1, 2, \cdots$ it holds $(T_{m+k,x} u)_s - (T_{k,x} u)_s = (Sx)_{m+k} - (Tx)_{m+k} + (T\mu_{m+k,x} u)_{s+m+k} - (Sx)_k + (Tx)_k - (T\mu_{k,x} u)_{s+k} = (Tx)_k$ $-(Tx)_{m+k} + (T\mu_{m,x}\mu_{k,x} u)_{s+m+k} - (T\mu_{k,x} u)_{s+k} = (Tx)_k - (Tx)_{m+k} + (Ty)_{s+m+k}$ $-(TS^m y)_{s+k} = (Tx)_k - (Tx)_{m+k} + (S^m(Ty))_{s+k} - (TS^m y)_{s+k} = 0$, because of (8). This completes the proof.

It is clear that if X is not necessarily linear, Theorem 2 states as follows:

Theorem 3. *A point (x, T) is m-periodic with respect to the semi-flow γ, if and only if it holds $S^m x = x$* and $\mathrm{TS^m u = S^m Tu}$, *for all $u \in D_T$ with $u_k = x_k$, $k =$ $0, 1, \cdots, m$.*

Corollary 1. *If a point $(x, T_{0,x})$ is periodic with period m, then the sequence Tx is also periodic with the same period, whenever it is bounded.*

[5] For continuous skew product semi-flows consult [48]. A method of skew products of dynamical systems, which is powerful to examine the geometrical structures of trajectories in dynamical systems, was first studied by Anzai [2] in connection with isomorphy problems in ergodic theory. See, also, [19].

Proof. From (6) we conclude that the quantity $(TS^m x)_{lm} - (S^m Tx)_{lm} = (Tx)_{lm} - (Tx)_{(l+1)m} =: \xi$ is fixed and it does not depend on $l = 0, 1, 2, \cdots$. Thus $(Tx)_m - (Tx)_{(l+1)m} = l\xi$, which due to the fact that Tx is bounded, implies $\xi = 0$. Now, the result follows from (8).

Corollary 2. *If a point $(x, T_{0,x})$ is periodic with period m and the sequence Tx is bounded, then it holds $S^m T = TS^m$.*

It is well known that a point of a dynamical or semi-dynamical system is a rest point if it is periodic with period any positive number. Therefore, a point (x, T) is a rest point if x is a constant sequence, say $x = a$ and (6) holds for any $m \in \mathbf{N}$.

Theorem 4. *Consider the Volterra type operator T defined by (4) and assume that the pair (x, T) is a periodic point with period m. Then the sequence x must be periodic with period m and, for a certain sequence (b_k) depending on x, the operator T can be written in the convolution form*

$$(Tu)_k = b_k + \sum_{j=0}^{k} h_{k-j}(u_j), \;\; k = 0, 1, 2, \cdots . \tag{9}$$

Proof. From Theorem 2 the sequence x must be periodic with period m, and moreover for all $n, k \in \mathbf{N}_0$ and $u \in D_T$, with $u_j = x_j$ for $j = 0, 1, \cdots m$ we have $a_0 + g_{0,0}(x_0) - a_m - \sum_{j=0}^{m} g_{m,j}(x_j) = a_k + \sum_{j=0}^{k} g_{k,j}(u_{j+m}) - a_{k+m} - \sum_{j=0}^{m+k} g_{k+m,j}$ $(u_j) = a_k + \sum_{j=0}^{k} g_{k,j}(u_{j+m}) - a_{k+m} - \sum_{j=0}^{m-1} g_{k+m,j}(c) - \sum_{j=0}^{k} g_{k+m,j+m}(u_{j+m})$. Thus the quantity $\sum_{j=0}^{k}[g_{k+m,j+m}(u_{j+m}) - g_{k,j}(u_{j+m})]$ does not depend on the quantity u. This means that we can write $g_{k+m,j+m}(\xi) - g_{k,j}(\xi) = h_{k,j,m}$, where the right side does not depend on ξ. Put $j = 0$ and $k + m = r$. Then we have $g_{r,m}(\xi) = g_{r-m,0}(\xi) + h_{r-m,0,m} =: h_{r-m}(\xi) + p_{r-m,m}$. It shows that the original operator T can be written in the form (9), where $b_k := a_k + \sum_{j=0}^{k} p_{k-j,j}$. This proves the result.

Example 1. Consider the Nemytski type discrete difference equation

$$v_{n+1} = g_n(v_n, v_{n-1}, \cdots, v_{n-\rho}), \;\; n = 0, 1, \cdots , \tag{10}$$

in the reals, where ρ is a positive integer. In order to write it in the form $Sx = Tx$, we use the idea of LaSalle [43] and work on the $\rho + 1$-dimensional space by defining $(Tx)_n := \big(g_n(x_n^1, x_n^2, \cdots, x_n^{\rho+1}), x_n^1, x_n^2, \cdots, x_n^{\rho}\big)^T$, where $x_n^1 := v_n$, $x_n^2 := v_{n-1}$, $\cdots, x_n^{\rho+1} := v_{n-\rho}$. This equation has the form $x_{n+1} = f_n(x_n)$ studied in terms of processes elsewhere, see, e.g., [41]. Any solution of equation $Sx = Tx$ has first coordinate a solution of the original scalar equation.

Theorem 5. *Let x be a sequence of points in $\boldsymbol{R}^{\rho+1}$. If the pair $(x, T_{0,x})$ is periodic with period $m(\geq \rho)$, then the sequences x and (g_n) are periodic with period m.*

Proof. From Theorem 2 we must have $S^m x = x$, and for any sequence u of points in $\mathbf{R}^{\rho+1}$, condition (6) must be satisfied. Then we have $g_k(u^1_{\rho+k}, u^2_{\rho+k}, \cdots, u^{\rho+1}_k) - g_{m+k}(u^1_{\rho+k}, u^2_{\rho+k}, \cdots, u^{\rho+1}_k) = g_0(x^1_0, x^2_0, \cdots, x^{\rho+1}_0) - g_m(x^1_m, x^2_m, \cdots, x^{\rho+1}_m) =: M$, for each $k = 0, 1, 2, \cdots$. Since the arguments $u^1_{\rho+k}, u^2_{\rho+k}, \cdots, u^{\rho+1}_k$ for $k \geq 0$ are arbitrary, we must have $g_{k+m}(\xi_0, \xi_1, \cdots, \xi_\rho) = g_k(\xi_0, \xi_1, \cdots, \xi_\rho) - M$, for all vectors in $\mathbf{R}^{m+1}$. Inductively, we can, easily, obtain $g_{\lambda m}(\xi_0, \xi_1, \cdots, \xi_\rho) = -\lambda M + g_0(\xi_0, \xi_1, \cdots, \xi_\rho)$, for all $\lambda = 0, 1, \cdots$. Since $(g_n(\xi_0, \xi_1, \cdots, \xi_\rho))$ forms a bounded sequence, we must have $M = 0$. This completes the proof.

Before giving some compactness conditions on the semi flow, we need to define a class of sequences as follows: Let $x \in E$ and (t_k) be a sequence of positive integers. The symbol $\mathscr{S}(x, (t_k))$ will denote the set of all two parameter sequences (s^{m,t_k}) such that $s^{m,t_k}_n = x_n$ for all $n = 0, 1, \cdots, t_k$ and all m, k as well as the limits $\lim_m S^{t_k} s^{m,t_k}$, $\lim_k S^{t_k} s^{m,t_k}$ exist and the successive limits $\lim_k \lim_m S^{t_k} s^{m,t_k}$, $\lim_m \lim_k S^{t_k} s^{m,t_k}$ exist and are equal. We start with the case of X being a linear metric space.

Theorem 6. *Let $x \in E$ with $\mathscr{R}(x)$ relatively compact. Also, assume that (1) for any sequence (t_k) of positive integers converging to $+\infty$, there is a subsequence (t_l) such that for any sequence (s^{m,t_l}) in $\mathscr{S}(x, (t_l))$, converging to some point with respect to m, the limit*

$$\lim_l [(Tx)_{t_l} - S^{t_l}(Ts^{m,t_l})] \tag{11}$$

exists for a certain m, and (2) the limit $\lim_m S^{t_l}(Ts^{m,t_l})$ exists uniformly with respect to l, in the l_∞ topology. Then the motion of the point $(x, T_{0,x})$ is compact.

Proof. Let (t_k) be a sequence of positive integers. If (t_k) is bounded, we can assume that it converges to some t, thus $t_k = t$, eventually. This shows that $S^{t_k}x = S^t x$, for all large k. For the second coordinate of the motion, we have $(T_{t_k,x}u)_k = (Sx)_{t_k} - (Tx)_{t_k} + (T\mu_{t_k,x}u)_k = (Sx)_t - (Tx)_t + (T\mu_{t,x}u)_k$, for all large k. Thus we see that the $(S^{t_k}x, T_{t_k,x})$ converges to $(S^t x, T_{t,x})$.

Now, assume that the sequence (t_k) is not bounded and it converges to $+\infty$. Also, due to Proposition 1(ii), we can assume that the sequence $(S^{t_k}x)$ converges to a certain limit $\bar{x}$. By assumption (1) there is a subsequence (t_l) of (t_k) such that (11) holds for every (s^{m,t_l}) in $\mathscr{S}(x, (t_l))$. Let (u^l) be a sequence in the domain of T converging to some $u \in E$. We set $s^{m,t_l} := \mu_{t_l,x}u^l$ for all m and l. Then observe that (s^{m,t_l}) is an element of $\mathscr{S}(x, (t_l))$. Therefore by (11) there is a sequence $a \in E$ such that $\lim_l T_{t_l,x}u^l = \lim_l [(Sx)_{t_l} - (Tx)_{t_l} + S^{t_l}(T(s^{m,t_l})] = a$. We claim that the limit a does not depend on the sequence (u^l). Indeed, let (v^l) be another sequence converging to u. Then, as above we conclude that there is some b such that $\lim_l T_{t_l,x}v^l = b$. Define a new sequence (w^l) as follows: $w^1 := u^1$, $w^2 := v^2$, $w^3 := u^3, \cdots$ which converges to u. Define $\bar{s}^{m,t_l} := \mu_{t_l,x}w^l$ and apply (8). Then the limit $\lim_l T_{t_l,x}w^l$ exists and is equal to a and b. Thus $a = b$. Therefore we can write $a = Ru$, for an operator

R. By the previous construction we have $R = \lim T_{t_l,x}$, and moreover R is causal.

It remains to show that R is continuous. To do that we let (u^m) be a sequence in E converging to some u^0. For any $m = 0, 1, \cdots$, we define the sequence $r_n^{m,t_l} := (1 - e^{-nl})u_n^m + e^{-nl}x_{t_l}$, $n = 0, 1, \cdots$ and let $s^{m,t_l} := \mu_{t_l,x}r^{m,t_l}$. Since x is bounded, we observe that $s_p^{m,t_l} = x_p$, for all $p < t_l$, $\lim_l \lim_m s^{m,t_l} = u^0 = \lim_m \lim_l s^{m,t_l}$, thus s^{m,t_l} it belongs to $\mathscr{S}(x, (t_l))$. Also, we have $\lim_m s^{m,t_l} = s^{0,t_l}$, thus from condition (2) and the continuity of T we conclude that

$$\lim_m S^{t_l}(Ts^{m,t_l}) = S^{t_l}(Ts^{0,t_l}) \tag{12}$$

exists uniformly with respect to l. Moreover, since and $\lim_l s^{m,l} = u^m$, for all m, and $\lim_l T_{t_l,x} = R$, we have

$$\lim_l T_{t_l,x}s^{m,t_l} = Ru^m. \tag{13}$$

Fix a $k \in \mathbf{N}_0$ and let $\epsilon > 0$. By (12) it follows that there is some m_0 such that $d((T_{t_l,x}s^{m,t_l})_k, (T_{t_l,x}s^{0,t_l})_k) < \epsilon$, for all $m \geq m_0$ and all indices l. Fix an $\bar{m} \geq m_0$. From (13) there is some index l_0 such that for all $l \geq l_0$ we have $d((T_{t_l,x}s^{\bar{m},t_l})_k, (Ru^{\bar{m}})_k) \leq \epsilon$ and $d((T_{t_l,x}s^{0,t_l})_k, (Ru^0)_k) \leq \epsilon$. Last three relations imply that $d((Ru^{\bar{m}})_k, (Ru^0)_k) \leq d((Ru^{\bar{m}})_k, (T_{t_l,x}s^{\bar{m},t_l})_k) + d((T_{t_l,x}s^{\bar{m},t_l})_k, (T_{t_l,x}s^{0,t_l})_k) + d((T_{t_l,x}s^{0,t_l})_k, (Ru^0)_k) < 3\epsilon$. The proof is complete.

If X is not necessarily linear, then the previous theorem states as follows and its proof is quite similar to the previous one:

Theorem 7. *Assume that the assumptions of Theorem 6 keep in force, except relation (11) which is replaced with the fact that the limit* $\lim_l S^{t_l}(Ts^{m,t_l})$ *exists for a certain* m*. Then the motion of the point* (x, T) *is compact.*

4 Full Limiting Orbits and Limiting Equations Along Solutions

Here we introduce the notion of the full limiting orbits of a pair (x, T), when x is a solution of equation $Sx = Tx$. (Notice that in this case we have $T_{0,x} = T$). These items are points of the form (x^*, T^*), where x^* is a full limiting sequence of x and T^* is an operator acting on the set of two-sided sequences. Also given any point $(\bar{x}, \bar{T})$ in the ω limit set $\omega(x, T)$ of (x, T), we construct a full orbit $(\bar{x}, \bar{T})$ in $\omega(x, T)$.

Consider an operator T, such that the pair (x, T) is compact, where x is a solution of $Sx = Tx$. Let $(\bar{x}, \bar{T})$ be a point in $\omega(x, T)$. Thus there is a sequence (t_n) of positive integers converging to $+\infty$, such that $\lim(S^{t_n}x, T_{t_n,x}) = (\bar{x}, \bar{T})$. By Proposition 2 there is a subsequence (r_n) of (t_n) which generates a full trajectory $U(\bar{x})$ of $\bar{x}$ and a full limiting sequence x^* such that $U(\bar{x})(0) = \bar{x}$ and $x^*_{k+j} = (S^kx^*)_j = (U(\bar{x})(k))_j = \lim(S^{t_n+k}x)_j = \lim_n x_{t_n+k+j}$, for all $j = 0, 1, \cdots$. Fix a $k \in \mathbf{Z}$. Then eventually $r_n +$

k is a nonnegative integer. By compactness there is a subsequence (s_n) of (r_n) with the property that $\lim_n T_{s_n+k,x} = \lim_n (T_{s_n,x})_{k,S^{t_n}x} =: T^*_{k,\bar{x}}$.

Proposition 4. *The two-sided sequence* $V(\bar{x},\bar{T})_k := Y_k := (U(\bar{x})(k), T^*_{k,\bar{x}}),\ \ k \in \mathbf{Z}$ *defines a full orbit through* $(\bar{x},\bar{T})$ *and it stays in* $\omega(x,T)$.

Proof. First we show $Y_k \in \omega(x,T)$, for all $k \in \mathbf{Z}$. Indeed, let k be fixed. There is a sequence (t_n) converging to $+\infty$, such that $U(\bar{x})(k)_j = \lim(S^{t_n+k}x)_j$. Also, for all $j \in \mathbf{N}_0$ we have $(T^*_{k,\bar{x}})_{j,U(\bar{x})(k)} = \lim_n (T_{t_n+k,x})_{j,S^{t_n+k}x} = \lim_n T_{t_n+k+j,x} = T^*_{k+j,\bar{x}}$. Therefore we have $\gamma(j;Y(k)) = \gamma(j;(U(\bar{x})(k), T^*_{k,\bar{x}}) = (S^jU(\bar{x})(k), (T^*_{k,\bar{x}})_{j,U(\bar{x})(k)})$ $= (U(\bar{x})(k+j), T^*_{k+j,\bar{x}})) = Y_{k+j}$. Notice that $Y_0 = (U(\bar{x})(0), T^*_{0,\bar{x}}) = (\bar{x},\bar{T})$. This completes the proof.

Consider a solution x of equation $Sx = Tx$. Then the translation of T along x at m is given by (5). Moreover we can easily observe that for any positive integer m it holds $S(S^m x) = T_{m,x}S^m x$, that is the sequence $(S^m x)$ solves the operator equation $u = T_{m,x}u$. Assume that the point (x,T) is compact, with respect to the semi-flow γ and let $(\bar{x},\bar{T}) \in \omega(x,T)$. Then for each $k \in \mathbf{Z}$ fixed, there is a sequence (t_n) converging to $+\infty$ such that $\lim_n(S^{t_n+k}x, T_{t_n+k,x}) = (U(\bar{x})(k), T^*_{k,\bar{x}})$. Also, due to the previous observation we get $S(S^{t_n+k}x) = T_{t_n+k,x}S^{t_n+k}x$, and keeping in mind the convergence structure of the space of operators, we get $S(U(\bar{x})(k)) = T^*_{k,\bar{x}}U(\bar{x})(k)$. Hence we proved the following result:

Theorem 8. *Let x be a solution of equation $Sx = Tx$, where the pair (x,T) is compact. Then any full limiting sequence of the solution x satisfies a full limiting equation of the original operator equation.*

5 Some Applications

Application 1. In [8] the local asymptotic stability of the positive solutions of equation $x_{n+1} = \beta + 1/x_n x_{n-1}$ was discussed, and by using a result from [7] it was proved that if $\beta \geq 4^{-1/3}$, then any positive solution has a finite limit. Here, by using the semi flow, we show the following more general result:

Theorem 9. *Consider the (scalar) difference equation $x_{n+1} = \beta_n + \frac{\alpha_n}{x_n x_{n-1}}$, where the sequences (α_n) and (β_n) have positive terms and converge to some positive α and β, respectively. If $\beta \geq (\alpha/4)^{1/3}$, then any positive solution converges to the unique positive root of equation $u^3 - \beta u^2 - \alpha = 0$.*

Proof. Any solution x of equation is bounded since for a fixed $\varepsilon \in (0,\beta)$ it holds $x_{n+1} \geq \beta - \varepsilon$ and $x_{n+1} \leq \beta + \varepsilon + (\alpha+\varepsilon)/(\beta-\varepsilon)^2$, eventually. Define the full limiting sequences y and z of the solution x such that $y_0 = \limsup x_n$ and $z_0 = \liminf x_n$. These two two-sided sequences satisfy the full limiting equation $u_{n+1} = \beta + \frac{\alpha}{u_n u_{n-1}},\ \ n \in \mathbf{Z}$. From this we get $y_0 \leq \beta + \alpha/z_0^2$ and $z_0 \geq \beta + \alpha/y_0^2$.

Combining these two relations we finally conclude that $y_0 = z_0$. This proves the convergence of the solution. The uniqueness of the (positive) equilibrium of equation, follows from the fact that the two curves $y = v - \beta$ and $y = a/v^2$ intersect at a unique point in the positive orthant.

Application 2. Here we prove the following result which extends a result in [41], Chapter 2:

Theorem 10. *Consider the difference equation* $x_{n+1} = \frac{\beta_n x_n + \gamma_n x_{n-1}}{A_n + B_n x_n + C_n x_{n-1}}$, $n = 0, 1, \cdots$, *where the sequences* (β_n), (γ_n), (A_n), (B_n), (c_n) *have nonnegative terms and converge to nonnegative reals* β, γ, A, B, C *respectively satisfying* $\beta + \gamma \leq A$, *with* $A > 0$. *Then any solution starting from nonnegative initial values converge to 0.*

Proof. Let (x_n) be a solution starting from nonnegative initial values. Then all terms of the sequence are nonnegative and satisfy $x_{n+1} < \frac{1}{B_n}(\beta_n - \gamma_n \frac{B_n}{C_n}) + \frac{\gamma_n}{C_n}$. The right side converges to a finite limit, which means that the solution is bounded. Let y and z be the full limiting sequences of (x_n) as in Theorem 9. The result will follow if we show that $y_0 = 0$. Indeed, we have

$$y_0 = \frac{\beta y_{-1} + \gamma y_{-2}}{A + B y_0 + C y_{-1}} \leq \frac{(\beta + \gamma) y_0}{A + (B + C) z_0}. \tag{14}$$

If equality holds, then $y_0 = y_{-1} = y_{-2}$ and thus $y_0 = \frac{\beta y_{-1} + \gamma y_{-2}}{A + B y_0 + C y_{-1}} = \frac{(\beta + \gamma) y_0}{A + (B + C) y_0}$, and so $A + (B + C) y_0 = \beta + \gamma \leq A$. Hence $y_0 = 0$. If in (14) the strict inequality holds, then we get $A + (B + C) z_0 < \beta + \gamma \leq A$, a contradiction. Thus $y_0 = 0$ and the proof is complete.

Application 3. Next we shall discuss the difference equation $x_{n+1} = p_n + \frac{x_{n-1}}{x_n^k}$, $n = 0, 1, \cdots$, where k is a positive real number. We shall prove the following result:

Theorem 11. *Assume that* (p_n) *is bounded and* $\liminf p_n > k^{1/k} \geq 1$. *If* (p_n) *is slowly varying, then any solution is also slowly varying. Moreover, if* (p_n) *converges to some* p, *then any solution converges to the unique positive root* ρ *of the algebraic equation* $\rho = p + \rho^{1-k}$.

Proof. Fix some a, b such that $a > p > b > 1$ and let n_0 be chosen with the property that $p_n \in [a, b]$ for all $n \geq n_0$. Let $\mu := a^{-k} < 1$. Then we can find $x_{m+n_0+1} \leq b/(1 - \mu) + \max\{x_{n_0}, x_{n_0} + 1\}$, for all $m = 1, 2, \cdots$. Hence, the solution x is bounded. Let w be a full liming sequence of x. Then, it is generated by a sequence of integers (k_n). This sequence, also, generates a full limiting sequence of (p_n), say p, which must be a constant. Hence w satisfies the full limiting equation

$$w_{n+1} = p + \frac{w_{n-1}}{w_n^k}. \tag{15}$$

We show that (w_n) is a constant sequence. In order to prove it we let $\sup w_n := W$. If there is some index m such that $W = w_m$, then we let $y_n := w_{n+m}$, for any

$n \in \mathbf{Z}$. If there is some sequence (m_n) converging to $+\infty$ such that $W = \lim w_{m_n} = \limsup w_n$, then we get a full limiting sequence (y_n) generated by (m_n). We do the same, when (m_n) converges to $-\infty$. In any case the two-sided sequence (y_n) satisfies the difference Eq. (15), for all n. Similarly, we obtain a full limiting sequence (z_n) such that $z_0 = \inf w_n$. Hence we have $y_0 \leq p + y_0/z_0^k$ and $z_0 \geq p + z_0/y_0^k$. From these relations we get $y_0 = z_0$, which clearly, proves the first result.

If the limit of (p_n) exists, then p is the unique full limiting sequence of (p_n), thus (w_n) is a constant sequence, ρ say, satisfying (15). This means that $\rho = p + \rho^{1-k}$, and the proof is complete.

The previous arguments extend some results given in [13] (with k = 1), in [41], Chapter 4, (where $p_n = p > 1$ and $k = 1$), as well as in [27] (with $p_n = p$).

Application 4. Consider the exponential type difference equation $x_{n+1} = a_n + b_n x_{n-1} \exp(-g(x_n, x_{n-1}, \cdots, x_{n-m}))$, $n = 0, 1, 2 \cdots$, where the sequences (a_n) and (b_n) have positive terms and g is a positive function. We shall show the following result:

Theorem 12. *Assume that the sequences* (a_n) *and* (b_n) *converge to* $a(> 0)$ *and* $b(> 0)$*, respectively, and the function* $g(\xi_1, \xi_2, \cdots, \xi_{m+1})$*, is increasing with respect to each variable, the function* $G(t) := g(t, t, \cdots, t)$ *converges to* $+\infty$*, it is increasing and its first derivative is nonincreasing. If* $b(1 + tG'(t)) < \exp(G(t))$*, for all* $t \geq 0$*, then there is a unique positive real* K *satisfying the equation* $K = a + bK \exp(-G(K))$ *and any positive solution converges to* K*.*

Proof. It is not hard to see that the function $\Phi(t) := t - bt \exp(-G(t))$ maps the interval $[0, +\infty)$ onto itself and it is strictly increasing. Thus the equation $\Phi(t) = a$ admits a unique solution K. Now, by using our assumptions, we can show that (x_n) is bounded. Let y and z be the full limiting sequences of x as in previous applications. Then we obtain the pair of inequalities $y_0 \leq a + by_0 e^{-G(z_0)}$, $z_0 \geq a + bz_0 \exp(-G(y_0))(\geq a > 0)$ which imply $y_0 = z_0$ and the proof is complete.

The previous result extends a result of [22], when $g(\xi_1, \cdots, \xi_{m+1}) = \xi_1$, $a_n = a$ and $b_n = b$. An interesting application of the previous result is given when $g(\xi_1, \cdots, \xi_{m+1}) := \xi_1^{r_1} \xi_2^{r_2} \cdots \xi_{m+1}^{r_{m+1}}$, where $r_1 + r_2 + \cdots r_{m+1} = 1$.

Application 5. If (x_n) is a sequence of real numbers, define its *limiting oscillation* by the type

$$\mathscr{O}(x) := \limsup x_n - \liminf x_n.$$

Theorem 13. *Consider the difference Eq. (1) given in the Introduction and assume that* (a_n) *has positive terms and it converges to some* $a \in (0, 1)$*. Then any solution satisfies* $\mathscr{O}(x) \leq \mathscr{O}(b)/(1 - a)$*. Moreover if* $b_n \to b$*, then* $x_n \to b/(1 - a)$*.*

Proof. First we can show that $(|x_n|)$ is bounded. Let y and z be the full limiting sequences of (x_n), as in the previous application. Then we obtain $y_0 \leq ay_0 + b$ and $z_0 \geq az_0 + \liminf b_n$. Hence, on one hand we get $y_0 - z_0 \leq \mathscr{O}(b)/(1 - a))$, on the other hand $b \leq (1 - a)z_0 \leq (1 - a)y_0 \leq b$. These relations imply the results.

Application 6. Finally, we have the difference equation $x_{n+1} = \alpha_n + \beta_n/x_n$

Theorem 14. *Assume that a sequence (α_n) has nonnegative terms and it converges to a positive real α. Also assume that (β_n) is a bounded sequence with nonnegative terms. Then for any positive solution of the difference equation it holds $\mathcal{O}(x) \leq \mathcal{O}(b)/a$. In particular, if (β_n) converges to some $\beta (\geq 0)$, then the solution x converges to $\frac{1}{2}[\alpha + \sqrt{\alpha^2 + 4\beta}]$.*

Proof. Let a, A be lower and upper bounds of the sequence (α_n) and B an upper bound of (β_n). Then we have $x_{n+1} \geq a$ and $x_{n+1} \leq A + B/a$, for each n. Hence, any positive solution is bounded. Take the full limiting sequences (y_n) and (z_n), as in Theorem 9. Then they satisfy $y_0 \leq \alpha + \limsup \beta_n/z_0$ and $z_0 \geq \alpha + \liminf \beta_n/y_0$. Combining these two relations we get the result.

Corollary 3. *Consider the general Fibonacci sequence $\zeta_{n+1} = \alpha_n\zeta_n + \beta_n\zeta_{n-1}$ with any (positive) initial values. If the sequences (α_n) and (β_n) converge to some positive reals α, β respectively, then it holds* $\lim \zeta_{n+1}/\zeta_n = \frac{1}{2}[\alpha + \sqrt{\alpha^2 + 4\beta}]$.

Proof. The sequence $x_n := \zeta_n/\zeta_{n-1}$ satisfies the equation in Theorem 14.

References

1. Abraham, R.L., Gardini, L., Mira, C.: Chaos in Discrete Dynamical Systems. Springer, New York (1977)
2. Anzai, H.: Ergodic skew product transformations on the torus. Osaka Math. J. **3**, 83–99 (1951)
3. Agarwal, R.P.: Difference Equations and Inequalities: Theory, Methods, and Applications. Monographs and Textbooks in Pure and Applied Mathematics, 2nd edn. Marcel Dekker Inc., New York (2000)
4. Agarwal, R.P., Bohner, M., Grace, S.R., O'Regan, D.: Discrete Oscillation Theory. Hindawi Publicing Corp., New York (2005)
5. Agarwal, R.P., Wong, P.J.Y.: Advanced Topics in Difference Equations. Kluwer Academic Publ., Dordrecht (1997)
6. Ravi, P.: Agarwal, Difference Equations and Inequalities, Pure and Applied Mathematics. Marcel Dekker, Inc., New York (2000)
7. Amleh, A.M., Camouzis, E., Ladas, G.: On the dynamics of a rational difference equations I. Int. J. Differ. Eqn. **3**(1), 1–35 (2008)
8. Anisimova, A.: On the second order rational difference equation $x_{n+1} = \beta + \frac{1}{x_n x_{n-1}}$, in difference equations. Discrete dynamical systems and applications. In: ICDEA, Barcelona, Spain (2012)
9. Bahi, J.M., Guyeux, C.: Discrete Dynamical Systems and Chaotic Machines, Theorey and Applications. CRP Press, A Chapman & Hall Book, Taylor and Francis Group (2013)
10. Cheban, D.N.: Asymptotically Almost Periodic Solutions of Differential Equations. Hindawi Publishing Corporation, New York (2009)
11. Crisci, M.R., Kolmanovski, V.B., Russo, E., Vecchio, A.: Boundedness of discrete Volterra equations. J. Math. Anal. Appl. **211**, 106–130 (1997)
12. Devaney, R.: An Introduction to Chaotic Dynamical Systems, 2nd edn. Addison-Wesley, Boston (1989)

13. Devault, R., Kocic, V.L., Stutson, D.: Global behavior of solutions of the nonlinear difference equation $x_{n+1} = p_n + \frac{x_{n-1}}{x_n}$. J. Diff. Eqn. Appl. **11**, 707–719 (2005)
14. Robert, W.: Easton, Geometric Methods for Discrete Dynamical Systems. Oxford Engineering Science Series, vol. 50. Oxford University Press, New York (1998)
15. Kin, E.: Skew products of dynamical systems. Trans. Am. Math. Soc. **166**, 27–43 (1972)
16. Elaydi, S.N.: An Introduction to Difference Equations. Springer, New York (1995)
17. Elaydi, S.N.: An Introduction to Difference Equations. Undergraduate Texts in Mathematics, 3rd edn. Springer, New York (2005)
18. Elaydi, S., Sacker, R.J.: Skew-product dynamical systems: applications to difference equations. Trinity University, Digital Commons@ Trinity 4, 1–22 (2004)
19. Elaydi, S.N., Kocic, V.L.: Global stability of a nonlinear Voltrerra difference system. Differ. Eqn. Dyn. Syst. **2**, 337–345 (1994)
20. Galor, O.: Discrete Dynamical Systems. Springer, Berlin (2007)
21. Holmgen, R.A.: A First Course in Discrete Dynamical Systems. Springer, New York (1994)
22. El-Metwally, E., Grove, E.A., Ladas, G., Levins, R., Radin, M.: On the difference equation, $x_{n+1} = \alpha + \beta x_{n-1} e^{-x_n}$. Nonlinear Anal. **47**, 4623–4634 (2001)
23. El-Metwally, H., Grove, E.A., Ladas, G.: A global convergence result with applications to periodic solutions. J. Math. Anal. Appl. **245**, 161–170 (2000)
24. Marotto, F.R.: Introduction to Mathematical Modeling Using Discrete Dynamical Systems. Thomson Brooks/Cole, Pacific Grove (2006)
25. Goldberg, S.: Introduction to Difference Equations. Dover Publ., New York (1986)
26. Gunawardena, J.: Cycle times and fixed points of min-max functions. In: Cohen, G., Quadrat, J.P. (eds.) 11th International Conference on Analysis and Optimization of Systems Discrete Event Systems. Lecture Notes in Control and Information Sciences, vol. 199 (1994). Springer, Heidelberg
27. Hamza, A.E., Morsy, A.: On the recursive sequence $x_{n+1} = \alpha + \frac{x_{n-1}}{x_n^k}$. Appl. Math. Lett. **22**, 91–95 (2009)
28. Karakostas, G.L.: Causal operators and topological dynamics. Ann. Mat. Pura Appl. **131**, 1–27 (1982)
29. Karakostas, G.L.: A discrete semi-flow in the space of sequences and study of convergence of sequences defined by a recured way. Mathematiki Epitheorisi **36**, 66–74 (1989)
30. Karakostas, G.L., Stević, S.: On a difference equation with min-max response. Int. J. Math. Math. Sci. **55**, 2915–2926 (2004)
31. Karakostas, G.L., Stević, S.: On the recursive sequence $x_{n+1} = A + \frac{f(x_n, x_{n-1}, \cdots, x_{n-k+1})}{x_{n-k}}$. Comm. Appl. Nonlinear Anal. **11**(3), 87–99 (2004)
32. Karakostas, G.L., Stević, S.: On the recursive sequence $x_{n+1} = a + \frac{x_{n-k}}{f(x_n, x_{n-1}, \cdots, x_{n_k+1})}$. Demonstr. Math. **XXXVIII**(3), 595–610 (2005)
33. Karakostas, G.L., Stević, S.: On the difference equation $x_{n+1} = Af(x_n) + f(x_{n-1})$. Appl. Anal. **83**(3), 309–323 (2003)
34. Karakostas, G.L., Stević, S.: Slowly varying solutions of the difference equation $x_{n+1} = f(x_n, \cdots, x_{n-k}) + g(n, x_n, \cdots, x_{n-k})$. J. Diff. Eqn. Appl. **10**(3), 249–255 (2004)
35. Karakostas, G.L.: Asymptotic 2-periodic difference equation with diagonally self-invertible responses. J. Diff. Eqn. Appl. **6**, 329–335 (2000)
36. Karakostas, G.L.: Convergence of a difference equation via the full limiting sequences method. Differ. Eqn. Dyn. Syst. **1**(4), 289–294 (1993)
37. Karakostas, G.L., Philos, Ch., Sficas, Y.: The dynamics of some discrete population models. Nonlinear Anal. **17**(11), 1069–1084 (1991)
38. Kloeden, P.E., Loretz, T., Yang, M.: Forward attractors in discrete time nonautonomous dynamical systems. In: Pinelas, S. et al. (eds.) Differential and Difference Equations with Applications, pp. 313–322. Springer, Cham (2016)
39. Kloeden, P.E., Rasmussen, M.: Nonautonomous Dynamical Systems. Mathematical Surveys and Monographs, vol. 176. AMS (2011)

40. Kocic, V.L., Ladas, G.: Global Behavior of Nonlinear Difference Equations of Higher Order with Applications. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht (1993)
41. Kulenović, M.R.S., Ladas, G.: Dynamics of Second Order Rational Difference Equations. Chapman and Hall/CRC, New York (2002)
42. LaSalle, J.P.: The Stability of Dynamical Systems. In: CBMS Regional Conference Series in Applied Mathematics, no. 25. SIAM, Philadelphia (1976)
43. LaSalle, J.P.: The Stability and Control of Discrete Processes. Springer, New York (1986)
44. Li, T.-Y., Yorke, J.A.: Period three implies chaos. AMS Monthly **82**(10), 985–992 (1975)
45. Miller, R.K.: Almost periodic differential equations as dynamical systems with applications to the existence of A.P. solutions. J. Diff. Eqn. **1**, 337–345 (1965)
46. Patula, W.T., Voulov, H.D.: On the oscillation and periodic character of a third order rational difference equation. Proc. Am. Math. Soc. **131**(3), 905–909 (2002)
47. Petersen, K.: Ergotic Theory. Cambridge University Press, New York (1983)
48. Sacker, R.J., Sell, G.R.: Lifting properties in skew product flows with applications to differential equations. Memoirs of the AMS 11(190) (1977)
49. Sell, G.R.: Nonautonomous differential equations and topological dynamics, Parts I and II. Trans. Amer. Math. Soc., **127** (I 967), 241-262, 263-283
50. Sibirsky, S.: Introduction to Topological Dynamics. Noordhoff International Publishing, Leyden (1975)
51. Veselý, M.: Construction of almost periodic sequences with given properties. Electron. J. Diff. Eqn. **2011**, 1–16 (2011)
52. Veselý, M.: Almost periodic sequences and functions with given values. Archivum Mathematicum (BRNO) **47**(126), 1–22 (2008)
53. Xhao, X.-Q.: Dynamical Systems in Population Biology. CMS. Springer, New York (2017)
54. Zhang, W.-B.: Discrete Dynamical Systems, Bifurcations and Chaos in Economics. Elsevier B. V. (2006)

Research of Four-Dimensional Dynamic Systems Describing Processes of Three'level Assimilation

Temur Chilachava, Sandra Pinelas, and George Pochkhua

Abstract In this work a new nonlinear mathematical model of process of three level assimilation which is described by four-dimensional dynamic systems is offered. In case of constancy of coefficients special points of the dynamic system are found. The conditions on constant coefficients for which it is possible to find special points with all four coordinates non-negative are determined. Introducing some dependence among coefficients of the system, two first integrals are derived, and the four-dimensional system is reduced to a two-dimensional one.

The sign-variable divergence theorem of a two-dimensional vector field in some one-coherent area of the first quadrant of the phase plane is proved. According to Bendixon's criterion it is possible to have a closed integral curve completely lying in this area.

Introduction. Mathematical modeling of physical processes has a long history. Mathematical modeling of physical processes involves the model adequacy, which is validated by Newton's non-relative five laws of classical mechanics: mass conservation law; law of conservation of impulse; the law of conservation the momentum of impulse; the first law of thermodynamics, i.e. energy conservation law; the second law of thermodynamics, i.e. entropy conservation law [1–5].

Creation of mathematical models is more original in social sphere, because, they are more difficult to substantiate [6–8].

We created a new direction of mathematical modeling, i.e. "Mathematical Modeling of Information Warfare" [9–11]. In these models two antagonistic sides waging with each other information warfare and also the third peacekeeping side trying to extinguish information warfare are considered. Conditions on model parameters

T. Chilachava (✉) · G. Pochkhua
Sokhumi State University, Sokhumi, Georgia
e-mail: temo_chilachava@yahoo.com

G. Pochkhua
e-mail: gia.pochkhua@gmail.com

S. Pinelas
Academia Militar, Amadora, Portugal
e-mail: sandra.pinelas@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_22

at which the third side will be able to force the conflicted sides to completion of information warfare are found.

We also offered mathematical models of forecasting the results of political elections in case of two or three parties. Also models in case of change of selective subjects before the next elections have been considered [12–16].

We proposed to create new nonlinear mathematical models of economic cooperation between two politically (not military opposition) mutually warring sides (two countries or a country and its legal region) which consider economic or other type of cooperation between different parts of population aimed to the peaceful resolution of conflicts [17, 18].

Taking into consideration the important tendencies in the world, it is important to study demographic and assimilation of social processes through mathematical modeling.

In [19] we considered a new nonlinear continuous mathematical model of linguistic globalization. Two categories of the world's population are considered: a category that hinders and a category leading to the dominant position of the English language. With a positive demographic factor of the population, which prevents globalization or a negative demographic factor of the population contributing to globalization, it is shown that the dynamic systems describing these processes allow the existence of two topologically not equivalent phase portraits (a stable node, a limit cycle).

It is known that, in the world, a social process of assimilation of languages is hidden. This process, as a rule, considers expansion of an area of the dominating languages (state languages of economically powerful states) at the expenses of less widespread languages (state languages economically of rather weak states).

According to this point of view, today, for less widespread languages (including classic languages) the conditions under which there will be no disappearance of the major languages are important, i.e. there will be no full assimilation of people talking in these languages.

1 System of the Equations and Initial Conditions

In this work a new nonlinear mathematical model of process of three-level assimilation which is described by four-dimensional dynamic system is offered:

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1(t)u(t) + \beta_1(t)u(t)v(t) + \beta_2(t)u(t)w(t) + \beta_3(t)u(t)z(t) \\ \frac{dv(t)}{dt} = \alpha_2(t)v(t) - \beta_4(t)u(t)v(t) + \beta_5(t)v(t)w(t) + \beta_6(t)v(t)z(t) \\ \frac{dw(t)}{dt} = \alpha_3(t)w(t) - \beta_7(t)u(t)w(t) - \beta_8(t)v(t)w(t) + \beta_9(t)w(t)z(t) \\ \frac{dz(t)}{dt} = \alpha_4(t)z(t) - \beta_{10}(t)u(t)z(t) - \beta_{11}(t)v(t)z(t) - \beta_{12}(t)w(t)z(t) \end{cases} \quad (1.1)$$

$$u(0) = u_0,\ v(0) = v_0,\ w(0) = w_0,\ z(0) = z_0, \quad (1.2)$$

$$\begin{array}{l} alpha_1(t) < 0,\ \alpha_4(t) > 0,\ \beta_i(t) > 0,\ i = \overline{1-12} \\ u,\ v,\ w,\ z \in C^1[0,T],\ t \in [0,T] \end{array} \tag{1.3}$$

$[0, T]$—assimilation process consideration period (as a rule, several decades, are possible till a century);
$u(t)$—the first population and powerful government institutions with very widespread language (**English**) influencing, through administrative resources, population talking in other three different languages;
$v(t)$—the second population and government institutions with widespread second language which undergoes assimilation from English, but in turn influencing, through administrative resources, the third and fourth populations with the purpose of their assimilation (for examples, French, German, Russian, Spanish, Chinese, Turkish and other);
$w(t)$—the third population which undergoes bilateral assimilation from two rather powerful states (for examples, Ukrainian, Arabic, Romanian (Moldavian), Catalan, and other);
$z(t)$—the fourth population which undergoes assimilation from the other three powerful languages (for examples, Occitan, Provencal language, Gagauz, and other) (Fig. 1);

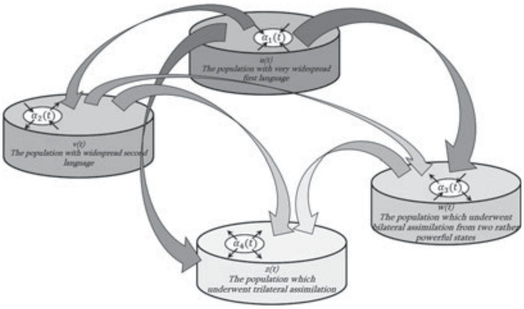

$\alpha_1(t)$
u(t)
The population with very widespread first language
$\alpha_2(t)$
v(t)
The population with widespread second language
$\alpha_3(t)$
w(t)
The population which underwent bilateral assimilation from two rather powerful states
$\alpha_4(t)$
z(t)
The population which underwent trilateral assimilation

Fig. 1 The scenario of process of three-level assimilation

The new mathematical model introduced in what follows assumes the natural inequalities:

$$\beta_i(t) > 0,\ i = \overline{1-12},\ t \in [0, T]. \tag{1.4}$$

We avoid to consider the trivial process of assimilation, in which a strong side completely assimilates the other three sides.

For the description of a nontrivial process, it is necessary to assume one among the following assumptions:

Assumption 1:

$$\begin{cases} \alpha_1(t) < 0 \\ \alpha_2(t) \geq 0 \\ \alpha_3(t) \geq 0 \\ \alpha_4(t) > 0 \end{cases} \quad t \in [0, T] \tag{1.5}$$

Assumption 2:

$$\begin{cases} \alpha_1(t) < 0 \\ \alpha_2(t) \leq 0 \\ \alpha_3(t) \leq 0 \\ \alpha_4(t) > 0 \end{cases} \quad t \in [0, T] \tag{1.6}$$

Assumption 3:

$$\begin{cases} \alpha_1(t) < 0 \\ \alpha_2(t) \leq 0 \\ \alpha_3(t) \geq 0 \\ \alpha_4(t) > 0 \end{cases} \quad t \in [0, T] \tag{1.7}$$

Assumption 4:

$$\begin{cases} \alpha_1(t) < 0 \\ \alpha_2(t) \geq 0 \\ \alpha_3(t) \leq 0 \\ \alpha_4(t) > 0 \end{cases} \quad t \in [0, T] \tag{1.8}$$

2 Some Special Cases

We will assume that all coefficients of system of the equations (1.1) are constants

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1 u(t) + \beta_1 u(t)v(t) + \beta_2 u(t)w(t) + \beta_3 u(t)z(t) \\ \frac{dv(t)}{dt} = \alpha_2 v(t) - \beta_4 u(t)v(t) + \beta_5 v(t)w(t) + \beta_6 v(t)z(t) \\ \frac{dw(t)}{dt} = \alpha_3 w(t) - \beta_7 u(t)w(t) - \beta_8 v(t)w(t) + \beta_9 w(t)z(t) \\ \frac{dz(t)}{dt} = \alpha_4 z(t) - \beta_{10} u(t)z(t) - \beta_{11} v(t)z(t) - \beta_{12} w(t)z(t) \end{cases} \tag{2.1}$$

$$u(0) = u_0,\ v(0) = v_0,\ w(0) = w_0,\ z(0) = z_0,$$

$$\beta_i > 0,\ i = \overline{1-12},\ \alpha_1 < 0,\ \alpha_4 > 0,\ t \in [0, T].$$

Stationary points of the nonlinear system of differential equations (2.1) are:

$$\begin{aligned} &M_0(0; 0; 0; 0),\ M_1(0; 0; \tfrac{\alpha_4}{\beta_{12}}; -\tfrac{\alpha_3}{\beta_9}),\ M_2(0; \tfrac{\alpha_4}{\beta_{11}}; 0; -\tfrac{\alpha_2}{\beta_6}), \\ &M_3(0; \tfrac{\alpha_3}{\beta_8}; -\tfrac{\alpha_2}{\beta_5}; 0),\ M_4(\tfrac{\alpha_4}{\beta_{10}}; 0; 0; -\tfrac{\alpha_1}{\beta_3}),\ M_5(\tfrac{\alpha_3}{\beta_7}; 0; -\tfrac{\alpha_1}{\beta_2}; 0), \\ &M_6(\tfrac{\alpha_2}{\beta_4}; -\tfrac{\alpha_1}{\beta_1}; 0; 0),\ M_7(0; v_*; w_*; z_*),\ M_8(u_{**}; 0; w_{**}; z_{**}), \\ &M_9(u_{***}; v_{***}; 0; z_{***}),\ M_{10}(u_{****}; v_{****}; w_{****}; 0), \\ &M_{11}(u_{*****}; v_{*****}; w_{*****}; z_{*****}), \end{aligned} \tag{2.2}$$

where

$$v_*,\, w_*,\, z_* : \begin{cases} \alpha_2 + \beta_5 w + \beta_6 z = 0 \\ \alpha_3 - \beta_8 v + \beta_9 z = 0 \\ \alpha_4 - \beta_{11} v - \beta_{12} w = 0, \end{cases}$$

$$u_{**},\, w_{**},\, z_{**} : \begin{cases} \alpha_1 + \beta_2 w + \beta_3 z = 0 \\ \alpha_3 - \beta_7 u + \beta_9 z = 0 \\ \alpha_4 - \beta_{10} u - \beta_{12} w = 0, \end{cases}$$

$$u_{***},\, v_{***},\, z_{***} : \begin{cases} \alpha_1 + \beta_1 v + \beta_3 z = 0 \\ \alpha_2 - \beta_4 u + \beta_6 z = 0 \\ \alpha_4 - \beta_{10} u - \beta_{11} v = 0, \end{cases}$$

$$u_{****},\, v_{****},\, w_{****} : \begin{cases} \alpha_1 + \beta_1 v + \beta_2 w = 0 \\ \alpha_2 - \beta_4 u + \beta_5 w = 0 \\ \alpha_3 - \beta_7 u - \beta_8 v = 0, \end{cases}$$

$$u_{*****},\, v_{*****},\, w_{*****},\, z_{*****} : \begin{cases} \alpha_1 + \beta_1 v + \beta_2 w + \beta_3 z = 0 \\ \alpha_2 - \beta_4 u + \beta_5 w + \beta_6 z = 0 \\ \alpha_3 - \beta_7 u - \beta_8 v + \beta_9 z = 0 \\ \alpha_4 - \beta_{10} u - \beta_{11} v - \beta_{12} w = 0. \end{cases}$$

We will enter transformation

$$\overline{u} = u - u_0,\ \overline{v} = v - v_0,\ \overline{w} = w - w_0,\ \overline{z} = z - z_0. \tag{2.3}$$

Then from (1.2), (2.1), (2.3) it is easy to receive

$$\begin{cases} \frac{\frac{d\overline{u}(t)}{dt}}{\overline{u}+u_0} = \alpha_1 + \beta_1 v_0 + \beta_2 w_0 + \beta_1 \overline{v} + \beta_2 \overline{w} + \beta_3 z_0 + \beta_3 \overline{z} \\ \frac{\frac{d\overline{v}(t)}{dt}}{\overline{v}+v_0} = \alpha_2 - \beta_4 u_0 + \beta_5 w_0 - \beta_4 \overline{u} + \beta_5 \overline{w} + \beta_6 z_0 + \beta_6 \overline{z} \\ \frac{\frac{d\overline{w}(t)}{dt}}{\overline{w}+w_0} = \alpha_3 - \beta_7 u_0 - \beta_8 v_0 - \beta_7 \overline{u} - \beta_8 \overline{v} + \beta_9 z_0 + \beta_9 \overline{z} \\ \frac{\frac{d\overline{z}(t)}{dt}}{\overline{z}+z_0} = \alpha_4 - \beta_{10} u_0 - \beta_{11} v_0 - \beta_{10} \overline{u} - \beta_{11} \overline{v} - \beta_{12} w_0 - \beta_{12} \overline{w} \end{cases} \tag{2.4}$$

We will pick up coefficients of a system of the Eq. (2.1) (model parameters) so that it was carried out

$$\begin{cases} \alpha_1 + \beta_1 v_0 + \beta_2 w_0 + \beta_3 z_0 = 0 \\ \alpha_2 - \beta_4 u_0 + \beta_5 w_0 + \beta_6 z_0 = 0 \\ \alpha_3 - \beta_7 u_0 - \beta_8 v_0 + \beta_9 z_0 = 0 \\ \alpha_4 - \beta_{10} u_0 - \beta_{11} v_0 - \beta_{12} w_0 = 0 \end{cases} \tag{2.5}$$

Existence of a set of positive solutions (u_0, v_0, w_0, z_0) of a system (2.5), requires that the determinants of the following five matrices are equal to zero:

$$\varDelta = \begin{vmatrix} 0 & \beta_1 & \beta_2 & \beta_3 \\ -\beta_4 & 0 & \beta_5 & \beta_6 \\ -\beta_7 & -\beta_8 & 0 & \beta_9 \\ -\beta_{10} & -\beta_{11} & -\beta_{12} & 0 \end{vmatrix} = 0 \tag{2.6}$$

$$\varDelta_1 = \begin{vmatrix} -\alpha_1 & \beta_1 & \beta_2 & \beta_3 \\ -\alpha_2 & 0 & \beta_5 & \beta_6 \\ -\alpha_3 & -\beta_8 & 0 & \beta_9 \\ -\alpha_4 & -\beta_{11} & -\beta_{12} & 0 \end{vmatrix} = 0 \tag{2.7}$$

$$\varDelta_2 = \begin{vmatrix} 0 & -\alpha_1 & \beta_2 & \beta_3 \\ -\beta_4 & -\alpha_2 & \beta_5 & \beta_6 \\ -\beta_7 & -\alpha_3 & 0 & \beta_9 \\ -\beta_{10} & -\alpha_4 & -\beta_{12} & 0 \end{vmatrix} = 0 \tag{2.8}$$

$$\varDelta_3 = \begin{vmatrix} 0 & \beta_1 & -\alpha_1 & \beta_3 \\ -\beta_4 & 0 & -\alpha_2 & \beta_6 \\ -\beta_7 & -\beta_8 & -\alpha_3 & \beta_9 \\ -\beta_{10} & -\beta_{11} & -\alpha_4 & 0 \end{vmatrix} = 0 \tag{2.9}$$

$$\varDelta_4 = \begin{vmatrix} 0 & \beta_1 & \beta_2 & -\alpha_1 \\ -\beta_4 & 0 & \beta_5 & -\alpha_2 \\ -\beta_7 & -\beta_8 & 0 & -\alpha_3 \\ -\beta_{10} & -\beta_{11} & -\beta_{12} & -\alpha_4 \end{vmatrix} = 0 \tag{2.10}$$

From (2.6)–(2.10) it is easy to receive

$$\begin{gathered}
\varDelta = \beta_4(\beta_3\beta_8\beta_{12} - \beta_2\beta_9\beta_{11} + \beta_1\beta_9\beta_{12}) - \beta_7(\beta_3\beta_5\beta_{11} - \beta_2\beta_6\beta_{11} + \beta_1\beta_6\beta_{12}) \\
+ \beta_{10}(\beta_1\beta_5\beta_9 - \beta_2\beta_6\beta_8 + \beta_3\beta_5\beta_8) = 0 \\
\\
\varDelta_1 = -\alpha_1(\beta_6\beta_8\beta_{12} - \beta_5\beta_9\beta_{11}) + \alpha_2(\beta_1\beta_9\beta_{12} - \beta_2\beta_9\beta_{11} + \beta_3\beta_8\beta_{12}) \\
- \alpha_3(\beta_3\beta_5\beta_{11} - \beta_2\beta_6\beta_{11} + \beta_1\beta_6\beta_{12}) + \alpha_4(\beta_1\beta_5\beta_9 - \beta_2\beta_6\beta_8 + \beta_3\beta_5\beta_8) = 0 \\
\\
\varDelta_2 = \alpha_1(\beta_6\beta_7\beta_{12} - \beta_5\beta_9\beta_{10} - \beta_4\beta_9\beta_{12}) - \alpha_2(\beta_3\beta_7\beta_{12} - \beta_2\beta_9\beta_{10}) \\
+ \alpha_3(\beta_3\beta_4\beta_{12} - \beta_2\beta_6\beta_{10} + \beta_3\beta_5\beta_{10}) + \alpha_4(\beta_2\beta_6\beta_7 - \beta_3\beta_5\beta_7 - \beta_2\beta_4\beta_9) = 0 \\
\\
\varDelta_3 = \alpha_1(-\beta_6\beta_7\beta_{11} + \beta_6\beta_8\beta_{10} + \beta_4\beta_9\beta_{11}) - \alpha_2(-\beta_3\beta_7\beta_{11} + \beta_1\beta_9\beta_{10} + \beta_3\beta_8\beta_{10}) \\
- \alpha_3(\beta_3\beta_4\beta_{11} - \beta_2\beta_6\beta_{10}) + \alpha_4(\beta_3\beta_4\beta_8 - \beta_1\beta_6\beta_7 + \beta_1\beta_4\beta_9) = 0 \\
\\
\varDelta_4 = \alpha_1(-\beta_4\beta_8\beta_{12} + \beta_5\beta_7\beta_{11} - \beta_5\beta_8\beta_{10}) - \alpha_2(\beta_2\beta_7\beta_{11} - \beta_2\beta_8\beta_{10} - \beta_1\beta_7\beta_{12}) \\
+ \alpha_3(\beta_2\beta_4\beta_{11} - \beta_1\beta_5\beta_{10} - \beta_1\beta_4\beta_{12}) - \alpha_4(\beta_2\beta_4\beta_8 - \beta_1\beta_5\beta_7) = 0
\end{gathered} \tag{2.11}$$

2.1 A First Reduction

Let's consider a special case, assuming the positions:

$$\begin{cases} \beta_1 = \beta_4 \\ \beta_2 = \beta_7 \\ \beta_3 = \beta_{10} \\ \beta_5 = \beta_8 \\ \beta_6 = \beta_{11} \\ \beta_9 = \beta_{12} \end{cases} \tag{2.12}$$

Then from (2.1), (2.12) it is easy to receive

$$\begin{aligned} &\Delta = (\beta_3\beta_5 - \beta_2\beta_6 + \beta_1\beta_9)^2 = 0 \\ &\Delta_1 = (\beta_3\beta_5 - \beta_2\beta_6 + \beta_1\beta_9)(\alpha_2\beta_9 - \alpha_3\beta_6 + \alpha_4\beta_5) = 0 \\ &\Delta_2 = (\beta_3\beta_5 - \beta_2\beta_6 + \beta_1\beta_9)(-\alpha_1\beta_9 + \alpha_3\beta_3 - \alpha_4\beta_2) = 0 \\ &\Delta_3 = (\beta_3\beta_5 - \beta_2\beta_6 + \beta_1\beta_9)(\alpha_1\beta_6 - \alpha_2\beta_3 + \alpha_4\beta_1) = 0 \\ &\Delta_4 = (\beta_3\beta_5 - \beta_2\beta_6 + \beta_1\beta_9)(-\alpha_1\beta_5 + \alpha_2\beta_2 - \alpha_3\beta_1) = 0 \end{aligned} \tag{2.13}$$

Thus, in this considered special case, we find

$$\beta_3\beta_5 + \beta_1\beta_9 = \beta_2\beta_6 \tag{2.14}$$

so that the following system holds:

$$\begin{cases} \Delta = 0 \\ \Delta_1 = 0 \\ \Delta_2 = 0 \\ \Delta_3 = 0 \\ \Delta_4 = 0. \end{cases} \tag{2.15}$$

Thus, from (2.4) and (2.5), (2.12) we will receive the following system

$$\begin{cases} \frac{\frac{d\overline{u}(t)}{dt}}{\overline{u}+u_0} = \beta_1\overline{v} + \beta_2\overline{w} + \beta_3\overline{z} \\ \frac{\frac{d\overline{v}(t)}{dt}}{\overline{v}+v_0} = -\beta_1\overline{u} + \beta_5\overline{w} + \beta_6\overline{z} \\ \frac{\frac{d\overline{w}(t)}{dt}}{\overline{w}+w_0} = -\beta_2\overline{u} - \beta_5\overline{v} + \beta_9\overline{z} \\ \frac{\frac{d\overline{z}(t)}{dt}}{\overline{z}+z_0} = -\beta_3\overline{u} - \beta_6\overline{v} - \beta_9\overline{w} \end{cases} \tag{2.16}$$

If we multiply the first Eq. (2.16) by γ_1, the second equation—by γ_2, the third equation—by γ_3, the fourth equation—by γ_4 and adding the obtained four equations of the system, then we get

$$\begin{aligned}&\tfrac{d}{dt}\left[\ln(\overline{u}+u_0)^{\gamma_1}(\overline{v}+v_0)^{\gamma_2}(\overline{w}+w_0)^{\gamma_3}(\overline{z}+z_0)^{\gamma_4}\right]\\&=-(\beta_1\gamma_2+\beta_2\gamma_3+\beta_3\gamma_4)\overline{u}+(\beta_1\gamma_1-\beta_5\gamma_3-\beta_6\gamma_4)\overline{v}\\&+(\beta_2\gamma_1+\beta_5\gamma_2-\beta_9\gamma_4)\overline{w}+(\beta_3\gamma_1+\beta_6\gamma_2+\beta_9\gamma_3)\overline{z}\end{aligned} \tag{2.17}$$

In (2.17) we will pick up $\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4$ so that the following algebraic linear system follows

$$\begin{cases}\beta_3\gamma_1+\beta_6\gamma_2+\beta_9\gamma_3=0\\\beta_2\gamma_1+\beta_5\gamma_2-\beta_9\gamma_4=0\\\beta_1\gamma_1-\beta_5\gamma_3-\beta_6\gamma_4=0\\\beta_1\gamma_2+\beta_2\gamma_3+\beta_3\gamma_4=0\end{cases} \tag{2.18}$$

The determinant of the system (2.18), owing to (2.13) must be zero

$$\Delta_*=\begin{vmatrix}\beta_3&\beta_6&\beta_9&0\\\beta_2&\beta_5&0&-\beta_9\\\beta_1&0&-\beta_5&-\beta_6\\0&\beta_1&\beta_2&\beta_3\end{vmatrix}=-\Delta=0. \tag{2.19}$$

Therefore, there is a set of nontrivial solutions $\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4$ of system (2.18). Taking into account (1.2), (2.3), (2.17), (2.18), a first integral of the system (2.16) takes the form

$$u^{\gamma_1}v^{\gamma_2}w^{\gamma_3}z^{\gamma_4}=u_0^{\gamma_1}v_0^{\gamma_2}w_0^{\gamma_3}z_0^{\gamma_4} \tag{2.20}$$

It is easy shown that owing to (2.13), (2.14), (2.19), the following property holds

$$rank\begin{pmatrix}\beta_3&\beta_6&\beta_9&0\\\beta_2&\beta_5&0&-\beta_9\\\beta_1&0&-\beta_5&-\beta_6\\0&\beta_1&\beta_2&\beta_3\end{pmatrix}=2 \tag{2.21}$$

therefore, in the system (2.18) two of the four unknown $\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4$, can be taken arbitrarily. In what follows, we take $\gamma_1=1,\ \gamma_4=1$, so that, deleting the first and fourth equation of the system (2.18), we find

$$\begin{cases}\gamma_1=1\\\gamma_2=\frac{\beta_9-\beta_2}{\beta_5}\\\gamma_3=\frac{\beta_1-\beta_6}{\beta_5}\\\gamma_4=1\end{cases} \tag{2.22}$$

Thus, the above first integral (2.20) becomes:

$$uv^{\frac{\beta_9-\beta_2}{\beta_5}}w^{\frac{\beta_1-\beta_6}{\beta_5}}z=u_0v_0^{\frac{\beta_9-\beta_2}{\beta_5}}w_0^{\frac{\beta_1-\beta_6}{\beta_5}}z_0 \tag{2.23}$$

2.2 A Second Reduction

Now we consider the special case when

$$\begin{cases} \beta_2 - \beta_9 = \beta_5 \\ \beta_6 - \beta_1 = \beta_5 \end{cases} \tag{2.24}$$

The system (2.24), taking into account equality (2.14), leads to the equation

$$\beta_3 = \beta_1 + \beta_2 \tag{2.25}$$

Thus, if the coefficients $\beta_i > 0,\ i = \overline{1-12}$, in the system (2.12) are connected by the equations

$$\begin{cases} \beta_1 = \beta_4 \\ \beta_2 = \beta_7 \\ \beta_3 = \beta_{10} \\ \beta_5 = \beta_8 \\ \beta_6 = \beta_{11} \\ \beta_9 = \beta_{12} \\ \beta_3\beta_5 + \beta_1\beta_9 = \beta_2\beta_6 \\ \beta_2 - \beta_9 = \beta_5 \\ \beta_6 - \beta_1 = \beta_5 \\ \beta_2 > \beta_5 \end{cases} \tag{2.26}$$

$$\begin{cases} \beta_1 = \beta_1 \\ \beta_2 = \beta_2 \\ \beta_3 = \beta_1 + \beta_2 \\ \beta_4 = \beta_1 \\ \beta_5 = \beta_5 \\ \beta_6 = \beta_1 + \beta_5 \\ \beta_7 = \beta_2 \\ \beta_8 = \beta_5 \\ \beta_9 = \beta_2 - \beta_5 \\ \beta_{10} = \beta_1 + \beta_2 \\ \beta_{11} = \beta_1 + \beta_5 \\ \beta_{12} = \beta_2 - \beta_5 \\ \beta_2 > \beta_5 \end{cases} \tag{2.27}$$

then the system (2.1) takes the form

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1 u(t) + \beta_1 u(t)v(t) + \beta_2 u(t)w(t) + (\beta_1 + \beta_2)u(t)z(t) \\ \frac{dv(t)}{dt} = \alpha_2 v(t) - \beta_1 u(t)v(t) + \beta_5 v(t)w(t) + (\beta_1 + \beta_5)v(t)z(t) \\ \frac{dw(t)}{dt} = \alpha_3 w(t) - \beta_2 u(t)w(t) - \beta_5 v(t)w(t) + (\beta_2 - \beta_5)w(t)z(t) \\ \frac{dz(t)}{dt} = \alpha_4 z(t) - (\beta_1 + \beta_2)u(t)z(t) - (\beta_1 + \beta_5)v(t)z(t) - (\beta_2 - \beta_5)w(t)z(t) \end{cases} \tag{2.28}$$

and its first integral becomes

$$uz = vwp, \qquad p \equiv \frac{u_0 z_0}{v_0 w_0} = const \tag{2.29}$$

According to Kronecker-Capelli's theorem, the non-homogeneous linear system (2.5) is compatible if and only if the rank of the system matrix is equal to that of the expanded matrix. By Eqs. (2.12), (2.14), (2.21) we have

$$rank \begin{pmatrix} 0 & \beta_1 & \beta_2 & \beta_3 \\ -\beta_1 & 0 & \beta_5 & \beta_6 \\ -\beta_2 & -\beta_5 & 0 & \beta_9 \\ -\beta_3 & -\beta_6 & -\beta_9 & 0 \end{pmatrix} = 2 \tag{2.30}$$

therefore, must be also two the rank of the expanded matrix:

$$rank \begin{pmatrix} 0 & \beta_1 & \beta_2 & \beta_3 & -\alpha_1 \\ -\beta_1 & 0 & \beta_5 & \beta_6 & -\alpha_2 \\ -\beta_2 & -\beta_5 & 0 & \beta_9 & -\alpha_3 \\ -\beta_3 & -\beta_6 & -\beta_9 & 0 & -\alpha_4 \end{pmatrix} = 2 \tag{2.31}$$

This leads to the following system for $\alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_4$

$$\begin{cases} \alpha_1\beta_5 - \alpha_2\beta_2 + \alpha_3\beta_1 = 0 \\ \alpha_2\beta_9 - \alpha_3\beta_6 + \alpha_4\beta_5 = 0 \\ \alpha_1\beta_6 - \alpha_2\beta_3 + \alpha_4\beta_1 = 0 \\ \alpha_1\beta_9 - \alpha_3\beta_3 + \alpha_4\beta_2 = 0 \end{cases} \tag{2.32}$$

Thus, under the conditions (2.27), (2.30)–(2.32), the system (2.5) for $u_0,\ v_0,\ w_0,\ z_0$, has a set of solutions and, as a same time, owing to (2.30), giving arbitrarily for example $u_0,\ v_0$, the other two unknowns $w_0,\ z_0$ are defined by the system (2.5).

From the system (2.32), owing to (2.27), it is easy to receive

$$\alpha_1 + \alpha_4 = \alpha_2 + \alpha_3 \tag{2.33}$$

Taking into account the assumptions of our model (not triviality of mathematical model), it is possible to assume additional restrictions for demographic factors $\alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_4$ of the sides

$$\begin{cases} \alpha_1 = \frac{\alpha_2\beta_2-\alpha_3\beta_1}{\beta_5} < 0 \\ \alpha_4 = \frac{\alpha_3\beta_6-\alpha_2\beta_9}{\beta_5} > 0 \end{cases} \tag{2.34}$$

For finding a second first integral of the system (2.28), we will consider its first three equations

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1 u(t) + \beta_1 u(t)v(t) + \beta_2 u(t)w(t) + (\beta_1+\beta_2)u(t)z(t) \\ \frac{dv(t)}{dt} = \alpha_2 v(t) - \beta_1 u(t)v(t) + \beta_5 v(t)w(t) + (\beta_1+\beta_5)v(t)z(t) \\ \frac{dw(t)}{dt} = \alpha_3 w(t) - \beta_2 u(t)w(t) - \beta_5 v(t)w(t) + (\beta_2-\beta_5)w(t)z(t) \end{cases} \tag{2.35}$$

After simple transformations in (2.28), we will receive

$$\begin{cases} \frac{1}{u}\frac{du}{dt} = \alpha_1 + \beta_1 v + \beta_2 w + (\beta_1+\beta_2)z \\ \frac{1}{v}\frac{dv}{dt} = \alpha_2 - \beta_1 u + \beta_5 w + (\beta_1+\beta_5)z \\ \frac{1}{w}\frac{dw}{dt} = \alpha_3 - \beta_2 u - \beta_5 v + (\beta_2-\beta_5)z \end{cases} \tag{2.36}$$

We will multiply the first equation of a system (2.36) on a, the second - on b, the third - on c, an the received equations we will put

$$\begin{aligned} &\frac{a}{u}\frac{du}{dt} + \frac{b}{v}\frac{dv}{dt} + \frac{c}{w}\frac{dw}{dt} = (\alpha_1 a + \alpha_2 b + \alpha_3 c) - (\beta_1 b + \beta_2 c)u + (\beta_1 a - \beta_5 c)v \\ &+(\beta_2 a + \beta_5 b)w + [(\beta_1+\beta_2)a + (\beta_1+\beta_5)b + (\beta_2-\beta_5)c]\, z \end{aligned} \tag{2.37}$$

Now we will pick up $a,\ b,\ c$ so that the system took place

$$\begin{cases} \alpha_1 a + \alpha_2 b + \alpha_3 c = 0 \\ \beta_1 b + \beta_2 c = 0 \\ \beta_1 a - \beta_5 c = 0 \\ \beta_2 a + \beta_5 b = 0 \\ (\beta_1+\beta_2)a + (\beta_1+\beta_5)b + (\beta_2-\beta_5)c = 0 \end{cases} \tag{2.38}$$

Owing to (2.32), the system (2.38) is equivalent to the following system

$$\begin{cases} \beta_1 b + \beta_2 c = 0 \\ \beta_1 a - \beta_5 c = 0 \\ \beta_2 a + \beta_5 b = 0 \end{cases} \tag{2.39}$$

The decision (2.39) has the following appearance

$$\begin{cases} a = a \\ b = -\frac{\beta_2}{\beta_5}a \\ c = \frac{\beta_1}{\beta_5}a \end{cases} \tag{2.40}$$

Taking into account (2.38), from (2.37) it is easy to receive the first integral of system (2.36)

$$u^a v^b w^c = u_0^a v_0^b w_0^c \tag{2.41}$$

For example, having taken $a = 1$, from (2.40) we will receive

$$\begin{cases} a = 1 \\ b = -\frac{\beta_2}{\beta_5} \\ c = \frac{\beta_1}{\beta_5} \end{cases}$$

and (2.41) will take a form

$$uv^{-\frac{\beta_2}{\beta_5}} w^{\frac{\beta_1}{\beta_5}} = u_0 v_0^{-\frac{\beta_2}{\beta_5}} w_0^{\frac{\beta_1}{\beta_5}} \tag{2.42}$$

If taking into account (2.27), we pick up

$$\begin{cases} \beta_1 = \beta_5 \\ \beta_2 = 2\beta_5 \\ \beta_5 = \beta_5 \end{cases} \tag{2.43}$$

then (2.42) will define the first integral of the system (2.36) or the second first integral of the system (2.28) taking into account (2.43)

$$\frac{uw}{v^2} = \frac{u_0 w_0}{v_0^2} \equiv q \tag{2.44}$$

Thus (2.29), (2.44), under the assumptions (2.34), (2.43), represent the two first integrals for the following system with initial conditions

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1 u(t) + \beta_1 u(t)v(t) + 2\beta_1 u(t)w(t) + 3\beta_1 u(t)z(t) \\ \frac{dv(t)}{dt} = \alpha_2 v(t) - \beta_1 u(t)v(t) + \beta_1 v(t)w(t) + 2\beta_1 v(t)z(t) \\ \frac{dw(t)}{dt} = \alpha_3 w(t) - 2\beta_1 u(t)w(t) - \beta_1 v(t)w(t) + \beta_1 w(t)z(t) \\ \frac{dz(t)}{dt} = \alpha_4 z(t) - 3\beta_1 u(t)z(t) - 2\beta_1 v(t)z(t) - \beta_1 w(t)z(t) \end{cases} \tag{2.45}$$

$$u(0) = u_0, v(0) = v_0, w(0) = w_0, z(0) = z_0$$

$$\alpha_1 = 2\alpha_2 - \alpha_3 < 0, \alpha_4 = 2\alpha_3 - \alpha_2 > 0$$

The two first integrals of the system (2.45) can be written in the following form:

$$\begin{cases} w = q\frac{v^2}{u} \\ z = pq\frac{v^3}{u^2} \end{cases} \tag{2.46}$$

Thus we have reduced the original four-dimensional dynamic system to the following two-dimensional one:

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1 u(t) + \beta_1 u(t)v(t) + 2\beta_1 q v^2(t) + 3\beta_1 pq \frac{v^3(t)}{u(t)} \\ \\ \frac{dv(t)}{dt} = \alpha_2 v(t) - \beta_1 u(t)v(t) + \beta_1 q \frac{v^3(t)}{u(t)} + 2\beta_1 pq \frac{v^4(t)}{u^2(t)} \end{cases} \tag{2.47}$$

$$u(0) = u_0,\, v(0) = v_0$$

Thus, we received a nonlinear two-dimensional dynamic system (2.47).

2.3 A Third Reduction

Now, let's us consider the special case when

$$\alpha_1 + \alpha_2 = 0 \tag{2.48}$$

which does not contradict conditions (2.33), (2.34) and leads to the system

$$\begin{cases} \alpha_1 < 0 \\ \alpha_2 = -\alpha_1 = |\alpha_1| > 0 \\ \alpha_3 = -3\alpha_1 = 3\,|\alpha_1| > 0 \\ \alpha_4 = -5\alpha_1 = 5\,|\alpha_1| > 0 \end{cases} \tag{2.49}$$

Thus, according to (2.47), (2.48), we get a two-dimensional dynamic system

$$\begin{cases} \frac{du(t)}{dt} = \alpha_1 u(t) + \beta_1 u(t)v(t) + 2\beta_1 q v^2(t) + 3\beta_1 pq \frac{v^3(t)}{u(t)} \\ \\ \frac{dv(t)}{dt} = -\alpha_1 v(t) - \beta_1 u(t)v(t) + \beta_1 q \frac{v^3(t)}{u(t)} + 2\beta_1 pq \frac{v^4(t)}{u^2(t)} \end{cases} \tag{2.50}$$

$$u(0) = u_0,\, v(0) = v_0$$

Theorem. The Cauchy problem (2.50), in some one-coherent area $D \subset (O, u(t), v(t))$ of the first quadrant of the phase plane $(O, u(t), v(t))$, has a closed integral curve completely lying in this area.

Proof. The system of nonlinear differential equations (2.50) will be written in vector form

$$\frac{d}{dt}\begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} F_1(u, v) \\ F_2(u, v) \end{pmatrix} \equiv \vec{F}, \tag{2.51}$$

where

$$\begin{aligned} F_1(u, v) &\equiv \alpha_1 u(t) + \beta_1 u(t)v(t) + 2\beta_1 q v^2(t) + 3\beta_1 pq \frac{v^3(t)}{u(t)} \\ F_2(u, v) &\equiv \alpha_2 v(t) - \beta_1 u(t)v(t) + \beta_1 q \frac{v^3(t)}{u(t)} + 2\beta_1 pq \frac{v^4(t)}{u^2(t)} \end{aligned} \tag{2.52}$$

Let's find the divergence of the two-dimensional vector field $\vec{F}$

$$\begin{array}{l} div\,\vec{F} = \nabla_i F^i = \frac{\partial F_1}{\partial u} + \frac{\partial F_2}{\partial v} \\ \frac{\partial F_1}{\partial u} = \alpha_1 + \beta_1 v - 3\beta_1 pq\frac{v^3}{u^2} \\ \frac{\partial F_2}{\partial v} = -\alpha_1 - \beta_1 u + 3\beta_1 q\frac{v^2}{u} + 8\beta_1 pq\frac{v^3}{u^2} \end{array} \tag{2.53}$$

Therefore

$$div\,\vec{F} = \beta_1 v - \beta_1 u + 3\beta_1 q\frac{v^2}{u} + 5\beta_1 pq\frac{v^3}{u^2} \tag{2.54}$$

We investigate the following function of two variables

$$\Phi(u, v) \equiv \beta_1 v - \beta_1 u + 3\beta_1 q\frac{v^2}{u} + 5\beta_1 pq\frac{v^3}{u^2} \tag{2.55}$$

and we will establish its signs in the first quarter of the phase plane (O, u, v).

By using Eq. (2.55), it is possible to show where, on some half-line (making physical sense for the model) of the first quadrant of the phase plane (O, u, v) the function of two variables $\Phi(u, v)$ vanishes.

We investigate the behavior of $\Phi(u, v)$ function on the (O, u, v) phase plane.

Let's consider a straight line

$$v = ku \tag{2.56}$$

Then, from (2.55), (2.56), we will receive

$$\Phi(u) = u\beta_1\left(k - 1 + 3k^2 q + 5pqk^3\right) \tag{2.57}$$

Let's enter designation

$$f(k) \equiv k - 1 + 3k^2 q + 5pqk^3 \tag{2.58}$$

It is easy to show that the equation

$$f(k) = 0 \tag{2.59}$$

has at least a positive root.

From (2.58) we find:

$$\begin{array}{l} f(0) = -1 < 0,\ f(1) = 3q + 5pq > 0,\ f(\infty) > 0,\ f(-\infty) < 0, \\ \frac{df}{dk} = 1 + 6qk + 15pqk^2 > 0,\ k > 0,\ f(k) \in C^\infty \end{array} \tag{2.60}$$

also there is at least a positive k_* such that

$$f(k_*) = 0$$

at the same time, by (2.60)

$$0 < k_* < 1, \ k_* \, \| u_0, v_0, w_0, z_0.$$

Thus the following conditions hold

$$\begin{array}{ll} v > k_* u & \Phi(u, v) > 0 \\ v < k_* u & \Phi(u, v) < 0 \\ v = k_* u & \Phi(u, v) = 0 \end{array} \tag{2.61}$$

Thus, given (2.54), (2.55), (2.61) it is shown that the divergence of the vector field in the first quadrant of the phase plane $(O, u(t), v(t))$ at the intersection of a straight line $v = k_* u$ changes its sign, hence according to the Bendixson's criterion, in some one-coherent area $D \subset (O, u(t), v(t))$ containing a segment of that straight line, it is possible to have a closed integral curve completely lying in this area.

Theorem is Proved

Under the conditions (2.27), (2.43), (2.49), the special point (2.2) of the nonlinear system of the differential equations (2.1), with all positive coordinates $M_{11}(u_{*****}; v_{*****}; w_{*****}; z_{*****})$, will be defined from the solution of the following linear algebraic system of equations $u_{*****}, \ v_{*****}, \ w_{*****}, \ z_{*****}$:

$$\begin{cases} v + 2w + 3z = -\frac{\alpha_1}{\beta_1} \\ -u + w + 2z = \frac{\alpha_1}{\beta_1} \\ 2u + v - z = -3\frac{\alpha_1}{\beta_1} \\ 3u + 2v + w = -5\frac{\alpha_1}{\beta_1} \end{cases} \tag{2.62}$$

The rank of the matrix of the system (2.62) is equal to two

$$rank \begin{pmatrix} 0 & 1 & 2 & 3 \\ -1 & 0 & 1 & 2 \\ 2 & 1 & 0 & -1 \\ 3 & 2 & 1 & 0 \end{pmatrix} = 2$$

Therefore two positive coordinates of a special point can be taken arbitrarily, and two others will be defined from two independent equations of the system (2.62)

$$\begin{aligned} w_{*****} &= -a\frac{\alpha_1}{\beta_1}, \\ z_{*****} &= -b\frac{\alpha_1}{\beta_1}, \\ u_{*****} &= -\frac{\alpha_1}{\beta_1}(1 + a + 2b), \\ v_{*****} &= -\frac{\alpha_1}{\beta_1}(1 - 2a - 3b). \end{aligned} \tag{2.63}$$

At the same time positive parameters need to be picked up so that the following system of inequalities hold

$$\begin{cases} a > 0 \\ b > 0 \\ (2 + k_*)a + (3 + 2k_*)b + k_* = 1 \end{cases} \tag{2.64}$$

If the third equation of the system (2.64) holds, we find

$$k_* = \frac{1 - 2a - 3b}{1 + a + 2b} \tag{2.65}$$

so that the point $M_{12}(u_{*****}; v_{*****})$ lies on the line $v = k_* u$ (O, u, v)phase plane, and it is also the projection of the point $M_{11}(u_{*****}; v_{*****}; w_{*****}; z_{*****})$ on to this phase plane.

Thus according to (2.55), (2.61) in the first quadrant of the phase plane $(O, u(t), v(t))$ there is such area in which $F(u, v)$ function of a sign change and according to Bendixson's criterion in this area existence of the closed integrated curve is possible, i.e. in this case according to (2.46) $w(t),\ z(t)$ functions doesn't become equal to zero and there is no full assimilation of the third and fourth languages.

Conclusion. Thus, assuming some relations between constant coefficients of the mathematical model, two first integrals of the nonlinear system of differential equations are obtained and the four-dimensional dynamic system is reduced to a two-dimensional one. The theorem of the sign-variable divergence of a two-dimensional vector field in some one-coherent area of the first quadrant of the phase plane has been proved. According to Bendixson 's criterion, a closed integral curve, that is, the existence of a non-zero solution, is possible in this area. For these constant model parameter values, no language is fully assimilated.

The authors express their deep gratitude to professors Paolo Ricci and Ilia Tavkhelidze for the repeated discussion of this work. Especially we want to note the useful advice of Professor Paolo Ricci, which contributed to improving the presentation of the results of the work.

References

1. Sedov, L.: Similarity and Dimensional Methods in Mechanics, 10th edn. CRC Press, Boca Raton (1993). 496 p
2. Golubyatnikov, A., Chilachava, T.: Estimates of the motion of detonation waves in a gravitating gas. Fluid Dyn. **19**(2), 292–296 (1984)
3. Chilachava, T.: Problem of a strong detonation in a uniformly compressing gravitating gas. Moscow State Univ. Bull. Ser. Math. Mech. **1**, 78–83 (1985)
4. Golubyatnikov, A., Chilachava, T.: Propagation of a detonation wave in a gravitating sphere with subsequent dispersion into a vacuum. Fluid Dyn. **21**(4), 673–677 (1986)
5. Chilachava, T.: A central explosion in an inhomogeneous sphere in equilibrium in its own gravitational field. Fluid Dyn. **23**(3), 472–477 (1988)
6. De Marchi, S.: Computational and Mathematical Modeling in the Social Sciences. Cambridge University Press, Cambridge (2005). 220 p
7. Reinhardt, H.E.: Mathematical Models in the Social Sciences (2012). 288 p

8. Strawinska-Zanko, U., Liebovitch, L.S.: Mathematical Modeling of Social Relationships: What Mathematics can tell us about People. Computational Social Sciences. Springer, Heidelberg (2018). 222 p
9. Chilachava, T., Kereselidze, N.: Optimizing problem of mathematical model of preventive information warfare. In: Information and Computer Technologies "Theory and Practice: Proceedings of the International Scientific Conference ICTMC–2010 Devoted to the 80th Anniversary of I.V. Prangishvili, USA, Imprint: Nova, pp. 525–529 (2012)
10. Chilachava, T., Kereselidze, N.: Mathematical modeling of information warfare. Inf. Warfare **1**(17), 28–35 (2011)
11. Chilachava, T., Chakhvadze, A.: Continuous nonlinear mathematical and computer model of information warfare with participation of authoritative interstate institutes. Georg. Electron. Sci. J.: Comput. Sci. Telecommun. **4**(44), 53–74 (2014)
12. Chilachava, T.: Nonlinear three-party mathematical model of elections, problems of management of safety of difficult systems. In: Works XXI of the International Conference, pp. 513–516 (2013)
13. Chilachava, T.: Nonlinear mathematical model of dynamics of voters of two political subjects. In: Seminar of I.Vekua Institute of Applied Mathematics, Reports, vol. 39, pp. 13–22 (2013)
14. Chilachava, T.: About some exact solutions of nonlinear system of the differential equations describing three-party elections. Appl. Math. Inf. Mech. **21**(1), 60–75 (2016)
15. Chilachava, T.: Mathematical model of transformation of two-party elections to three- party elections. Georg. Electron. Sci. J.: Comput. Sci. Telecommun. **2**(52), 21–29 (2017)
16. Chilachava, T., Sulava, L.: Mathematical and computer modeling of political elections. In: Proceedings of the Eleventh International Scientific–Practical Conference Internet-Education Science-2018, pp. 113–116 (2018)
17. Chilachava, T., Pochkhua, G.: About a possibility of resolution of conflict by means of economic cooperation. Problems of management of safety of difficult systems. In: In: The XXVI International Conference, Moscow, pp. 69–74 (2018)
18. Chilachava, T., Pochkhua, G.: Research of the nonlinear dynamic system describing mathematical model of settlement of the conflicts by means of economic cooperation. In: 8th International Conference on Applied Analysis and Mathematical Modeling, ICAAMM 2019, Proceedings Book, pp. 183–187 (2019)
19. Chilachava, T.: Research of the dynamic system describing globalization process. In: Proceedings in Mathematics & Statistics, Mathematics, Informatics and their Applications in Natural Sciences and Engineering, vol. 276, pp. 67–78. Springer (2019)

Two Positive Solutions for a Nonlinear Neumann Problem Involving the Discrete p-Laplacian

G. Bonanno, P. Candito, and G. D'Aguì

Abstract This paper is devoted to study of existence of at least two positive solutions for a nonlinear Neumann boundary value problem involving the discrete p-Laplacian.

Keywords Multiple solutions · Difference equations · Neumann problem

2010 Mathematics Subject Classification 39A10 · 39A12 · 34B15

1 Introduction

In this paper, we investigate the existence of two positive solutions for the following nonlinear discrete Neumann boundary value problem

$$\begin{cases} -\Delta(\phi_p(\Delta u(k-1))) + q(k)\phi_p(u(k)) = \lambda f(k, u(k)), & k \in [1, N], \\ \Delta u(0) = \Delta u(N) = 0, \end{cases} \qquad N_{\lambda,\underline{f}}$$

where λ is a positive parameter, N is a fixed positive integer, $[0, N+1]$ is the discrete interval $\{0, ..., N+1\}$, $\phi_p(s) := |s|^{p-2}s$, $1 < p < +\infty$ and for all $k \in [0, N+1]$, $q(k) > 0$, $\Delta u(k) := u(k+1) - u(k)$ denotes the forward difference operator and $f : [0, N+1] \times \mathbb{R} \to \mathbb{R}$ is a continuous function.

G. Bonanno · G. D'Aguì (✉)
Department of Engineering, University of Messina, Contrada Di Dio (S. Agata),
98166 Messina, Italy
e-mail: dagui@unime.it

G. Bonanno
e-mail: bonanno@unime.it

P. Candito
Department DICEAM, University of Reggio Calabria, Via Graziella (Feo Di Vito),
89122 Reggio Calabria, Italy
e-mail: pasquale.candito@unirc.it

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_23

The theory of difference equations employs numerical analysis, fixed point methods, upper an lower solutions methods (see, for instance, [3, 5, 7, 23]). The variational approach represents an important advance as it allows to prove multiplicity results, usually, under a suitable condition on the nonlinearities, see [1, 2, 7–11, 14–22, 24, 25].

In the present paper, we study the problem $(N_{\lambda,\underline{f}})$ following a variational approach, based on a recent result of Bonanno and D'Aguì (see [6]), that assures the existence of at least two non trivial critical points for a certain class of functionals defined on infinite-dimensional Banach space. This theorem is obtained by combining a local minimum result given in [13], together with the Ambrosetti-Rabinowitz theorem (see [4]). In the application of the mountain pass theorem, to prove the Palais-Smale condition of the energy functional associated to the nonlinear differential problems, the Ambrosetti-Rabinowitz condition is requested on the nonlinear term, in particular this means that the nonlinear term has to be more than p-superlinear at infinity.

In this paper, exploiting that the variational framework of the problem $(N_{\lambda,\underline{f}})$ is defined in a finite-dimensional space, we prove that the p-superlinearity at infinity of the primitive on the nonlinearity is enough to prove the Palais-Smale condition. For a complete overview on variational methods on finite Banach spaces and discrete problems, see [12]. We obtain, here, Theorem 2, which gived the existence of two positive solutions, by requiring an algebraic condition on the nonlinearity (we mean (6) in 2).

The paper is so organized: Sect. 2, contains basic definitions and main results on difference equations and some critical point tools, in addition, Lemma 2 is given in order to prove the Palais-Smale condition of the functional associated to problem $(N_{\lambda,\underline{f}})$. Section 3 is devoted to our main result. In particular, our main theorem allows us to obtain two positive solutions with only one hypothesis on the primitive of the nonlinear term f without any asymptotic behaviour at zero. Moreover, a consequence (Corollary 1) (requiring the p-superlinearity at infinity and the p-sublinearity at zero on the primitive of f) of our main result is presented in order to show the applicability of our results.

2 Mathematical Background

In the $N+2$-dimensional Banach space

$$X = \{u : [0, N+1] \to \mathbb{R} : \Delta u(0) = \Delta u(N) = 0\},$$

we consider the norm

$$\|u\| := \left(\sum_{k=1}^{N+1} |\Delta u(k-1)|^p + \sum_{k=1}^{N} q(k)|u(k)|^p\right)^{1/p} \quad \forall u \in X.$$

Moreover, we will use also the equivalent norm

$$\|u\|_\infty := \max_{k\in[0,N+1]} |u(k)|, \quad \forall u \in X.$$

For our purpose, it will be useful the following inequality

$$\|u\|_\infty \leq \|u\| q^{-1/p}, \quad \forall u \in X, \quad \text{where} \quad q := \min_{k\in[1,N]} q_k. \tag{1}$$

Moreover, we mention the classical Hölder norm on X.

$$\|u\|_p = \left(\sum_{k=0}^{N+1} |u(k)|^p \right)^{\frac{1}{p}}.$$

We observe that being X a finite dimensional Banach space, all norms defined on it are equivalent and in particular, there exist two positive constants L_1 and L_2 such that

$$L_1\|u\|_p \leq \|u\| \leq L_2\|u\|_p. \tag{2}$$

To describe the variational framework of problem $(N_{\lambda,\underline{f}})$, we introduce the following two functions

$$\Phi(u) := \frac{\|u\|^p}{p} \quad \text{and} \quad \Psi(u) := \sum_{k=1}^{N} F(k, u(k)), \quad \forall u \in X, \tag{3}$$

where $F(k,t) := \int_0^t f(k,\xi)d\xi$ for every $(k,t) \in [1,N] \times \mathbb{R}$. Clearly, Φ and Ψ are two functionals of class $C^1(X,\mathbb{R})$ whose Gâteaux derivatives at the point $u \in X$ are given by

$$\Phi'(u)(v) = \sum_{k=1}^{N+1} \phi_p\left(\Delta u\left(k-1\right)\right) \Delta v\left(k-1\right) + q(k)\left|u\left(k\right)\right|^{p-2} u\left(k\right) v\left(k\right),$$

and

$$\Psi'(u)(v) = \sum_{k=1}^{N} f\left(k, u\left(k\right)\right) v(k),$$

for all $u, v \in X$. Taking into account that

$$-\sum_{k=1}^{N} \Delta(\phi_p(\Delta u(k-1)))v(k) = \sum_{k=1}^{N+1} \phi_p(\Delta u(k-1))\Delta v(k-1), \quad \forall\, u\ \ v, \in X,$$

it is easy to verify, see also [25], that

Lemma 1. *A vector $u \in X$ is a solution of problem $(N_{\lambda,\underline{f}})$ if and only if u is a critical point of the function $I_\lambda = \Phi - \lambda\Psi$.*

Let $(X, \|\cdot\|)$ be a Banach space and let $I \in C^1(X, \mathbb{R})$. We say that I satisfies the Palais-Smale condition (in short (PS)-condition), if any sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ such that

1. $\{I(u_n)\}_{n\in\mathbb{N}}$ is bounded,
2. $\{I'(u_n)\}_{n\in\mathbb{N}}$ converges to 0 in X^*,

admits a subsequence which is convergent in X.

Here, we recall the abstract result established in [6], on the existence of two non-zero critical points.

Theorem 1. *Let X be a real Banach space and let $\Phi, \Psi : X \to \mathbb{R}$ be two functionals of class C^1 such that $\inf_X \Phi = \Phi(0) = \Psi(0) = 0$. Assume that there are $r \in \mathbb{R}$ and $\tilde{u} \in X$, with $0 < \Phi(\tilde{u}) < r$, such that*

$$\frac{\sup_{u\in\Phi^{-1}(]-\infty,r])} \Psi(u)}{r} < \frac{\Psi(\tilde{u})}{\Phi(\tilde{u})}, \tag{4}$$

and, for each

$$\lambda \in \Lambda = \left]\frac{\Phi(\tilde{u})}{\Psi(\tilde{u})}, \frac{r}{\sup_{u\in\Phi^{-1}(]-\infty,r])} \Psi(u)}\right[,$$

the functional $I_\lambda = \Phi - \lambda\Psi$ satisfies the (PS)-condition and it is unbounded from below.

Then, for each $\lambda \in \Lambda$, the functional I_λ admits at least two non-zero critical points $u_{\lambda,1}$, $u_{\lambda,2}$ such that $I(u_{\lambda,1}) < 0 < I(u_{\lambda,2})$.

Here and in the sequel we suppose $f(k, 0) \geq 0$ for all $k \in [1, N]$. We assume that $f(k, x) = f(k, 0)$ for all $x < 0$ and for all $k \in [1, N]$. Put

$$L_\infty(k) := \liminf_{s\to+\infty} \frac{F(k,s)}{s^p}, \quad L_\infty := \min_{k\in[1,N]} L_\infty(k).$$

We give the following lemma.

Lemma 2. *If $L_\infty > 0$ then I_λ satisfies (PS)-condition and it is unbounded from below for all $\lambda \in \left]\frac{L_2^p}{pL_\infty}, +\infty\right[$, where L_2 is given in (2).*

Proof. Since $L_\infty > 0$ we put $\lambda > \frac{L_2^p}{pL_\infty}$ and l such that $L_\infty > l > \frac{L_2^p}{p\lambda}$. Let $\{u_n\}$ be a sequence such that $\lim_{n\to+\infty} I_\lambda(u_n) = c$ and $\lim_{n\to+\infty} I'_\lambda(u_n) = 0$. Put $u_n^+ = \max\{u_n, 0\}$ and $u_n^- = \max\{-u_n, 0\}$ for all $n \in \mathbb{N}$. We have that $\{u_n^-\}$ is bounded. In fact, one has

$$\left|\Delta u_n^-(k-1)\right|^p \leq -\phi_p\left(\Delta u_n(k-1)\right)\Delta u_n^-(k-1),$$

for all $k \in [1, N+1]$, and

$$q(k)\left|u_n^-(k)\right|^p = -q(k)\left|u_n(k)\right|^{p-2} u_n(k)u_n^-(k),$$

for all $k \in [1, N+1]$.
So we have,

$$\begin{aligned}
&\sum_{k=1}^{N+1}\left(\left|\Delta u_n^-(k-1)\right|^p + q(k)\left|u_n^-(k)\right|^p\right)\\
&\leq -\sum_{k=1}^{N+1}\left(\phi_p\left(\Delta u_n(k-1)\right)\Delta u_n^-(k-1) + q(k)\left|u_n(k)\right|^{p-2} u_n(k)u_n^-(k)\right).
\end{aligned}$$

So,

$$\begin{aligned}
\|u_n^-\|^p &= \sum_{k=1}^{N+1}\left(\left|\Delta u_n^-(k-1)\right|^p + q(k)\left|u_n^-(k)\right|^p\right)\\
&\leq -\sum_{k=1}^{N+1}\left(\phi_p\left(\Delta u_n(k-1)\right)\Delta u_n^-(k-1) + q(k)\left|u_n(k)\right|^{p-2} u_n(k)u_n^-(k)\right)\\
&= -\Phi'(u_n)(u_n^-).
\end{aligned}$$

By definition of u_n^- and taking into account that $f(k,x) = f(k,0)$ for all $x < 0$ and for all $k \in [1, N]$, we have

$$\Psi'(u_n)(u_n^-) = \sum_{k=1}^{N} f\left(k, u_n(k)\right)u_n^-(k) \geq 0.$$

So, we get

$$\|u_n^-\|^p \leq -\Phi'(u_n)(u_n^-) \leq -\Phi'(u_n)(u_n^-) + \lambda\Psi'(u_n)(u_n^-),$$

that is

$$\|u_n^-\|^p \leq -I'_\lambda(u_n)(u_n^-), \tag{5}$$

for all $n \in \mathbb{N}$. Now, from $\lim\limits_{n\to+\infty} I'_\lambda(u_n) = 0$, one has $\lim\limits_{n\to+\infty} \dfrac{I'_\lambda(u_n)(u_n^-)}{\|u_n^-\|} = 0$, for which, taking (5) into account, gives $\lim\limits_{n\to+\infty} \|u_n^-\| = 0$. So, we obtain the claim. And, there is $M > 0$ such that $\|u_n^-\| \leq M$, $\|u_n^-\|_p \leq \dfrac{M}{L_1} = L$, $0 \leq u_n^-(k) \leq L$ for all $k \in [1, N]$ for all $n \in \mathbb{N}$.

At this point, by contradiction argument, assume that $\{u_n\}$ is unbounded (that is, $\{u_n^+\}$ is unbounded).

From $\liminf\limits_{s\to+\infty} \dfrac{F(k,s)}{s^p} = L_\infty(k) \geq L_\infty > l$ there is $\delta_k > 0$ such that $F(k,s) > ls^p$ for all $s > \delta_k$. Moreover,

$$\begin{aligned} F(k,s) &\geq \min_{s\in[-L,\delta_k]} F(k,s) \geq ls^p - l\,(\max\{\delta_k, L\})^p + \min_{s\in[-L,\delta_k]} F(k,s) \\ &\geq ls^p - \max\{l\,(\max \delta_k, L)^p - \min_{s\in[-L,\delta_k]} F(k,s), 0\} = ls^p - Q(k) \end{aligned}$$

for all $s \in [-L, \delta_k]$. Hence, $F(k,s) \geq ls^p - Q(k)$ for all $s \geq -L$. It follows that $F\,(k, u_n(k)) \geq l\,(u_n(k))^p - Q(k)$ for all $n \in \mathbb{N}$ and for all $k \in [1, N]$, $\sum\limits_{k=1}^{N} F(k,$ $u_n(k)) \geq \sum\limits_{k=1}^{N} \left[l\,(u_n(k))^p - Q(k)\right] = l\|u_n\|_p^p - \sum\limits_{k=1}^{N} Q(k) = l\|u_n\|_p^p - \overline{Q}$, that is,

$$\Psi(u_n) \geq l\|u_n\|_p^p - \overline{Q},$$

for all $n \in \mathbb{N}$. Therefore, one has

$$I_\lambda(u_n) = \Phi(u_n) - \lambda\Psi(u_n) = \frac{1}{p}\|u_n\|^p - \lambda\Psi(u_n) \leq \frac{L_2^p}{p}\|u_n\|_p^p - \lambda l\|u_n\|_p^p + \lambda\overline{Q},$$

that is

$$I_\lambda(u_n) \leq \left(\frac{L_2^p}{p} - \lambda l\right)\|u_n\|_p^p + \lambda\overline{Q},$$

for all $n \in \mathbb{N}$. Since $\|u_n\|_p \to +\infty$ and $\dfrac{L_2^p}{p} - \lambda l < 0$, one has $\lim\limits_{n\to+\infty} I_\lambda(u_n) = -\infty$ and this is absurd. Hence, I_λ satisfies (PS)-condition.

Finally, we get that I_λ is unbounded from below. Let $\{u_n\}$ be such that $\{u_n^-\}$ is bounded and $\{u_n^+\}$ is unbounded. As before, we obtain $\Psi(u_n) \geq l\|u_n\|_p^p - \overline{Q}$, for all

$n \in \mathbb{N}$ and, consequently, $I_\lambda(u_n) \leq \left(\frac{L_2^p}{p} - \lambda l\right) \|u_n\|_p^p + \lambda \overline{Q}$, for all $n \in \mathbb{N}$. Hence, $\lim\limits_{n \to +\infty} I_\lambda(u_n) = -\infty$ and the proof is complete.

3 Main Results

In this section, we present the main existence result of our paper. We start putting

$$Q = \sum_{k=1}^{N} q(k).$$

Theorem 2. *Let $f : [1, N] \times \mathbb{R} \to \mathbb{R}$ be a continuous function such that $f(k, 0) \geq 0$ for all $k \in [1, N]$, and $f(k, 0) \neq 0$ for some $k \in [1, N]$. Assume also that there exist two positive constants c and d with $d < c$ such that*

$$\frac{\sum\limits_{k=1}^{N} \max\limits_{|\xi| \leq c} F(k, \xi)}{c^p} < q \min \left\{ \frac{1}{Q} \frac{\sum\limits_{k=1}^{N} F(k, d)}{d^p}, \frac{L_\infty}{L_2^p} \right\}. \tag{6}$$

Then, for each $\lambda \in \bar{\Lambda}$ with

$$\bar{\Lambda} = \left] \max \left\{ \frac{Q}{p} \frac{d^p}{\sum\limits_{k=1}^{N} F(k, d)}, \frac{L_2^p}{p L_\infty} \right\}, \frac{q}{p} \frac{c^p}{\sum\limits_{k=1}^{N} \max\limits_{|\xi| \leq c} F(k, \xi)} \right[,$$

the problem $(N_{\lambda, \underline{f}})$ admits at least two positive solutions.

Proof. We consider the functionals Φ and Ψ given in (3). Φ and Ψ satisfy all regularity assumptions requested in Theorem 1, moreover we have that any critical point in X of the functional I_λ is exactly a solution of problem $(N_{\lambda, \underline{f}})$. Furthermore, $\inf\limits_S \Phi = \Phi(0) = \Psi(0) = 0$. In order to prove our result, we need to verify condition (4) of Theorem 1. Fix $\lambda \in \bar{\Lambda}$, from (6) one has that $L_\infty > 0$ and $\bar{\Lambda}$ is non-degenerate. From Lemma 2, the functional I_λ satisfies the (PS)-condition for each $\lambda > \frac{L_2^p}{p L_\infty}$, and it is unbounded from below. Now, put $r = \dfrac{qc^p}{p}$, an condier $u \in \Phi^{-1}\,(]-\infty, r])$; so such a u satisfies

$$\frac{1}{p}\|u\|^p \leq r,$$

so

$$\|u\| \leq (pr)^{\frac{1}{p}}.$$

One has

$$|u| \leq \frac{1}{q^{\frac{1}{p}}}\|u\| \leq \left(\frac{pr}{q}\right)^{\frac{1}{p}} = c.$$

So,

$$\Psi(u) = \sum_{k=1}^{N} F(k, u(k)) \leq \sum_{k=1}^{N} \max_{|\xi| \leq c} F(k, \xi),$$

for all $u \in X$ such that $u \in \Phi^{-1}\,(]-\infty, r])$.
Hence,

$$\frac{\sup\limits_{u \in \Phi^{-1}(]-\infty, r])} \Psi(u)}{r} \leq \frac{p}{q}\,\frac{\sum\limits_{k=1}^{N} \max\limits_{|\xi| \leq c} F(k, \xi)}{c^p}. \tag{7}$$

Now, let be $\tilde{u} \in \mathbb{R}^{N+2}$ be such that $\tilde{u}(k) = d$ for all $k \in [0, N+1]$. Clearly, $\tilde{u} \in X$ and it holds

$$\Phi(\tilde{u}) = \frac{Qd^p}{p}, \tag{8}$$

and so, we have

$$\frac{\Psi(\tilde{u})}{\Phi(\tilde{u})} = \frac{p}{Q}\,\frac{\sum\limits_{k=1}^{N} F(k, d)}{d^p}. \tag{9}$$

Therefore, from (7), (9) and assumption (6) one has

$$\frac{\sup\limits_{u \in \Phi^{-1}(]-\infty, r])} \Psi(u)}{r} < \frac{\Psi(\tilde{u})}{\Phi(\tilde{u})}.$$

Moreover, taking into account that $0 < d < c$ and again by (6), we have that

$$0 < d < \left(\frac{q}{Q}\right)^{\frac{1}{p}} c. \tag{10}$$

Indeed, by contradiction, if we suppose that $d \geq \left(\frac{q}{Q}\right)^{\frac{1}{p}} c$, we have

$$\frac{\sum_{k=1}^{N} \max_{|\xi|\leq c} F(k,\xi)}{c^p} \geq \frac{\sum_{k=1}^{N} F(k,d)}{c^p} \geq \frac{q}{Q}\frac{\sum_{k=1}^{N} F(k,d)}{d^p},$$

which contradicts (6). Hence by (8) and (10) we get $0 < \Phi(\tilde{u}) < r$.

So, finally we obtain that I_λ admits at least two non-zero critical points and then, for all $\lambda \in \bar{\Lambda} \subset \Lambda$, these are non zero solutions of $(N_{\lambda,\underline{f}})$.

Since we are interested to obtain a positive solution for problem $(N_{\lambda,\underline{f}})$, we adopt the following truncation on the functions $f(k,s)$,

$$f^+(k,s) = \begin{cases} f(k,s), \text{ if } s \geq 0; \\ f(k,0), \text{ if } s < 0. \end{cases}$$

Fixed $\lambda \in \Lambda_c^+$. Working with the truncations $f^+(k,s)$, since we have that $f(k(0,s) \neq 0$ for some $k \in [1,N]$, let u a non trivial solution guaranteed in the first part of the proof, now, to prove the u is nonnegative, we exploit the u is a critical point of the energy functional $I_\lambda = \Phi - \lambda\Psi$ associated to problem (N_{λ,f^+}). In other words, we have that $u \in X$ satisfies the following condition

$$\sum_{k=1}^{N+1} \phi_p(\Delta u(k-1))\Delta v(k-1) + \sum_{k=1}^{N} q(k)\phi_p(u(k))v(k) = \sum_{k=1}^{N} f^+(k,u(k))v(k), \ \forall u,v \in X. \tag{11}$$

From this, taking as test function $v = -u^-$, it is a simple computation to prove that $\|u^-\| = 0$, that is u is nonnegative. Moreover, arguing by contradiction, we show that u is also a positive solution of problem $(N_{\lambda,\underline{f}})$. Suppose that $u(k) = 0$ for some $k \in [1,N]$. Being u a solution of problem $(N_{\lambda,\underline{f}})$ we have

$$\phi_p(\Delta u(k-1)) - \phi_p(\Delta u(k)) = f(k,0) \geq 0,$$

which implies that

$$0 \geq -|u(k-1)|^{p-2}u(k-1) - |u(k+1)|^{p-2}u(k+1) \geq 0.$$

So, we have that $u(k-1) = u(k+1) = 0$. Hence, iterating this process, we get that $u(k) = 0$ for every $k \in [1, N]$, which contradicts that u is nontrivial and this completes the proof.

Now, we present a particular case of Theorem 2.

Corollary 1. *Assume that f is a continuous function such that $f(k, 0) > 0$ for all $k \in [0, N]$ and*

$$\limsup_{t\to 0^+} \frac{F(k,t)}{t^p} = +\infty, \tag{12}$$

and

$$\lim_{t\to +\infty} \frac{F(k,t)}{t^p} = +\infty,$$

for all $k \in [0, N]$, and put $\lambda^* = \dfrac{q}{p} \sup_{c>0} \dfrac{c^p}{\sum_{k=1}^{N} \max_{|\xi|\le c} F(k,\xi)}$.

Then, for each $\lambda \in]0, \lambda^*[$, the problem $(N_{\lambda,\underline{f}})$ admits at least two positive solutions.

Proof. First, note that $L_\infty = +\infty$. Then, fix $\lambda \in]0, \lambda^*[$ and $c > 0$ such that

$$\lambda < \frac{q}{p} \frac{c^p}{\sum_{k=1}^{N} \max_{|\xi|\le c} F(k,\xi)}.$$

From (12) we have

$$\limsup_{t\to 0^+} \frac{\sum_{k=1}^{N} F(k,t)}{t^p} = +\infty,$$

then there is $d > 0$ with $d < c$ such that $\dfrac{p}{Q} \dfrac{\sum_{k=1}^{N} F(k,d)}{d^p} > \dfrac{1}{\lambda}$. Hence, Theorem 2 ensures the conclusion.

Acknowledgements The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). The paper is partially supported by PRIN 2017 - Progetti di Ricerca di rilevante Interesse Nazionale, "Nonlinear Differential Problems via Variational, Topological and Set-valued Methods" (2017AYM8XW).

References

1. Agarwal, R.P., Perera, K., O'Regan, D.: Multiple positive solutions of singular and nonsingular discrete problems via variational methods. Nonlinear Anal. **58**, 69–73 (2004)
2. Agarwal, R.P., Perera, K., O'Regan, D.: Multiple positive solutions of singular discrete p-Laplacian problems via variational methods. Adv. Diff. Equ. **2**, 93–99 (2005)
3. Agarwal, R.P.: On multipoint boundary value problems for discrete equations. J. Math. Anal. Appl. **96**(2), 520–534 (1983)
4. Ambrosetti, A., Rabinowitz, P.H.: Dual variational methods in critical point theory and applications. J. Funct. Anal. **14**, 349–381 (1973)
5. Anderson, D.R., Rachuånková, I., Tisdell, C.C.: Solvability of discrete Neumann boundary value probles. Adv. Diff. Equ. **2**, 93–99 (2005)
6. Bonanno, G., D'Aguì, G.: Two non-zero solutions for elliptic Dirichlet problems. Z. Anal. Anwend. **35**(4), 449–464 (2016)
7. C. Bereanu, J. Mawhin, *Boundary value problems for second-order nonlinear difference equations with discrete ϕ-Laplacian and singular ϕ*, J. Difference Equ. Appl. **14** (2008), 1099–1118
8. C. Bereanu, P. Jebelean, C. Şerban, *Ground state and mountain pass solutions for discrete $p(\cdot)-$Laplacian*, Bound. Value Probl. 2012 (104) (2012)
9. Bonanno, G., Candito, P.: Nonlinear difference equations investigated via critical point methods. Nonlinear Anal. **70**, 3180–3186 (2009)
10. Bonanno, G., Candito, P.: Infinitely many solutions for a class of discrete nonlinear boundary value problems. Appl. Anal. **88**, 605–616 (2009)
11. Bonanno, G., Candito, P.: Nonlinear difference equations through variational methods, Handbook on Nonconvex Analysis, pp. 1–44. Int. Press, Somerville, MA (2010)
12. Bonanno, G., Candito, P., D'Aguì, G.: Variational methods on finite dimensional Banach spaces and discrete problems. Advanced Nonlinear Studies **14**, 915–939 (2014)
13. Bonanno, G.: A critical point theorem via the Ekeland variational principle. Nonlinear Anal. **75**, 2992–3007 (2012)
14. Bonanno, G., Jebelean, P., Şerban, C.: Superlinear discrete problems. Appl. Math. Lett. **52**, 162–168 (2016)
15. P. Candito, G. D'Aguì, *Three solutions for a discrete nonlinear Neumann problem involving the p-Laplacian*, Adv. Difference Equ. 2010, Art. ID 862016, 11 pp
16. Candito, P., D'Aguì, G.: Three solutions to a perturbed nonlinear discrete Dirichlet problem. J. Math. Anal. Appl. **375**, 594–601 (2011)
17. Candito, P., D'Aguì, G.: Constant-sign solutions for a nonlinear neumann problem involving the discrete p-laplacian. Opuscula Math. **34**(4), 683–690 (2014)
18. Candito, P., Giovannelli, N.: Multiple solutions for a discrete boundary value problem involving the p-Laplacian. Comput. Math. Appl. **56**, 959–964 (2008)
19. G. D'Aguì, J. Mawhin, A. Sciammetta *Positive solutions for a discrete two point nonlinear boundary value problem with p-Laplacian*, J. Math. Anal.Appl. **447**, (2017),383–397
20. M. Galewski, S. Głąb, *On the discrete boundary value problem for anisotropic equation*, J. Math. Anal. Appl. **386** (2012), 956–965
21. A. Guiro, I. Nyanquini, S. Ouaro, *On the solvability of discrete nonlinear Neumann problems involving the p(x)-Laplacian*, Adv. Difference Equ. 2011 (32) (2011)
22. L. Jiang, Z. Zhou, *Three solutions to Dirichlet boundary value problems for p-Laplacian difference equations*, Adv. Difference Equ. (2008). Article ID 345916, 10 p
23. Rachuånková, I., Tisdell, C.C.: Existence of non-spurious solutions to discrete Dirichlet problems with lower and upper solutions. Nonlinear Anal. **67**, 1236–1245 (2007)
24. Şerban, C.: Existence of solutions for discrete p-Laplacian with potential boundary conditions. J. Difference Equ. Appl. **19**, 527–537 (2013)
25. Tian, Y., Ge, W.: The existence of solutions for a second-order discrete Neumann problem with p-Laplacian. J. Appl. Math. Comput. **26**, 333–340 (2008)

Structure of Solution Sets for Fractional Partial Integro-Differential Equations

Hedia Benaouda

Abstract Our aim in this paper is to study in the first part the existence of mild solutions to the following partial fractional integro-differential equation with nonlocal conditions and in the second one we deal with extending the classical Kneser's theorem and Aronszajn type result for this class of equations by showing that the set of all solutions is a compact and R_δ-set.

1 Introduction

In this work, we study the existence of mild solutions and topological structure of the solution set to the following partial integro-differential equation with nonlocal conditions:

$$ {}^cD^\alpha x(t) = Ax(t) + \int_0^t B(t-s)x(s)ds + f(t, x(t)), \quad t \in [0, b], \tag{1}$$

$$ x(0) = x_0 + h(x), \tag{2}$$

where ${}^cD^\alpha$ stand for the Caputo fractional derivative, $0 < \alpha \leq 1$, $A : D(A) \subset E \to E$ is a closed linear operator on a Banach space $(E, |.|)$, $(B(t))_{t\geq 0}$ is a family of closed linear operators on E having the same domain $D(B) \supset D(A)$ which is independent of t, $f : [0, b] \times E \to E$ and $h : C([0, b], E) \to E$ are given functions satisfying conditions to be specified later, $C([0, b], E)$ stands for the space of continuous functions from $[0, b]$ to E endowed with the uniform norm topology. Equations of the form (1)–(2) serve as an abstract formulation of many partial integrodifferential equations arising in heat flow in materials with memory, viscoelasticity and many other physical phenomena [2]. The problem of existence and uniqueness of partial differential equations have been studied by many authors using different approaches [3–7]. The rest of this paper is organized as follows. In Sect. 2, we give some preliminary results on the fractional calculus. In Sect. 3, we study the existence of mild solutions

H. Benaouda (✉)
University Ibn Khaldoun Tiaret, Tiaret, Algeria
e-mail: b_hedia@univ-tiaret.dz

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_24

for (1)–(2) under some hypothesis using Mönch fixed point theorem. In Sect. 4 we extend the classical Kneser's theorem and Aronszajn type result by proving that the set of all solutions is compact and R_δ-set.

2 Preliminaries

In this section, we introduce some facts about t the Caputo fractional derivative that are used throughout this paper. Let $J := [0, b]$, $b > 0$ and $(E, |\cdot|)$ be a Banach space. $C(J, E)$ be the space of E-valued continuous functions on J endowed with the uniform norm topology

$$\|x\|_\infty = \sup\{\|x(t),\ t \in J\}.$$

$L^1(J, E)$ the space of E-valued Bochner integrable functions on J with the norm

$$\|f\|_{L^1} = \int_0^b \|f(t)\|dt.$$

Definition 1 ([10, 11]). The fractional (arbitrary) order integral of the function $h \in L^1([a, b], R_+)$ of order $\alpha \in R_+$ is defined by

$$I_a^\alpha h(t) = \int_a^t \frac{(t-s)^{\alpha-1}}{\Gamma(\alpha)} h(s)ds,$$

where Γ is the gamma function. When $a = 0$, we write $I^\alpha h(t) = h(t) * \varphi_\alpha(t)$, where $\varphi_\alpha(t) = \dfrac{t^{\alpha-1}}{\Gamma(\alpha)}$ for $t > 0$, and $\varphi_\alpha(t) = 0$ for $t \leq 0$, and $\varphi_\alpha \to \delta(t)$ as $\alpha \to 0$, where δ is the delta function.

Definition 2 ([10, 11]). For a function h given on the interval $[a, b]$, the Caputo fractional-order derivative of h of order $\alpha \in R_+$, is defined by

$$({}^cD_{a+}^\alpha h)(t) = \frac{1}{\Gamma(1-\alpha)} \int_a^t (t-s)^{-\alpha} h^{(n+1)}(s)ds,$$

where $n = [\alpha] + 1$.

Let us recall the following definitions and results that will be used in the sequel.

Definition 3. Let E be a real Banach space and $(Y, \leq)$ a partially ordered set. A function $\beta : \mathscr{P}(E) \to Y$ is called a measure of noncompactness in E if

$$\beta(\Omega) = \beta(\overline{\text{co}}\Omega)$$

for every $\Omega \subset \mathscr{P}(E)$, where $\overline{\text{co}}\Omega$ denotes the closed convex hull of Ω.

Definition 4. [8] A measure of noncompactness β is called:

(i) monotone if $\Omega_0, \Omega_1 \in \mathscr{P}(E)$, $\Omega_0 \subset \Omega_1$ implies $\beta(\Omega_0) \leq \beta(\Omega_1)$
(ii) nonsingular if $\beta(\{a\} \cup \Omega) = \beta(\Omega)$ for every $a \in E$, $\Omega \in \mathscr{P}(E)$;
(iii) invariant with respect to union with compact sets, if $\beta(\{K\} \cup \Omega) = \beta(\Omega)$ for every $K \in \mathscr{P}_k(E)$ and $\Omega \in \mathscr{P}(E)$.

If Y is a cone in a normed space, we say that the MNC is

(iv) regular if $\beta(\Omega) = 0$ is equivalent to the relative compactness of Ω.
(v) algebraically semiadditive, if $\beta(\Omega_0 + \Omega_1) \leq \beta(\Omega_0) + \beta(\Omega_1)$ for each Ω_0, $\Omega_1 \in \mathscr{P}(E)$.

One of most important example of a measure of noncompactness possessing all these properties is the Kuratowski measure of noncompactness defined by:

$$\nu(X) := \inf \left\{ \epsilon > 0 \ : \ \ X \subseteq \bigcup_{i=1}^{n} B_i \ \ and \ \ diam(B_i) \leq \epsilon \right\}.$$

where $diam(B_i) = \sup\{\|x - y\|; \ \ x, y \in B_i\}$.

Lemma 1. [10, 11] *If $\{u_n\}_{n=1}^{+\infty} \subset L^1(J, E)$ satisfies $\|u_n(t)\| \leq \kappa(t)$ a.e. on J for all $n \geq 1$ with some $\kappa \in L^1(J, R_+)$. Then the function $\nu(\{u_n(t)\}_{n=1}^{+\infty})$ belongs to $L^1(J, R_+)$ and*

$$\nu\left(\left\{\int_0^t u_n(s)ds \ : \ n \geq 1\right\}\right) \leq 2\int_0^t \nu(u_n(s)ds \ : \ n \geq 1)ds. \tag{3}$$

Lemma 2. *[1] Let E be a Banach space, $C \subset E$ be closed and bounded and $F : C \to E$ a condensing map. Then $I - F$ is proper and $I - F$ maps closed subsets of C onto closed sets.*

Recall that the map $I - F$ is proper if it is continuous and for every compact $K \subset E$, the set $(I - F)^{-1}(K)$ is compact. The application of the topological degree theory for condensing maps implies the following fixed point principle.

Theorem 1. *(Mönch fixed point theorem) [9] Let D be a bounded, closed and convex subset of a Banach X space such that* $0 \in D$*, and let T be a continuous mapping of D into itself. If the implication*

$$V = co\bar{n}vN(V), \qquad or \qquad V = N(V) \cup \{0\} \Rightarrow \nu(V) = 0$$

holds for every subset V of D, then N has a fixed point.

Definition 5. [1] Let $A \subset \mathscr{P}(X)$. The set A is called a contractible space provided there exists a continuous homotopy $H : A \times [0, 1] \to A$ and $x_0 \in A$ such that

(a) $H(x, 0) = x$, for every $x \in A$,
(b) $H(x, 1) = x_0$, for every $x \in A$,

i.e. if the identity map is homotopic to a constant map (A is homotopically equivalent to a point). Note that if $A \in \mathscr{P}_{cv,cl}(X)$, then A is contractible, but the class of contractible sets is much larger than the class of closed convex sets.

Definition 6. $A \in \mathscr{P}(X)$ is a retract of X if there exists a continuous map $r : X \to A$ such that

$$r(a) = a, \quad \text{for every } \ a \in A.$$

Definition 7. A compact nonempty space X is called an R_δ–set provided there exists a decreasing sequence of compact nonempty contractible spaces $\{X_n\}_{n=1}^{+\infty}$ such that $X = \bigcap_{n=1}^{+\infty} X_n$.

Let us recall the well-known Lasota-Yorke approximation lemma.

Lemma 3. [1] *Let E be a normed space, X a metric space and $F : X \to E$ be a continuous map. Then, for each $\varepsilon > 0$, there is a locally Lipschitz map $F_\varepsilon : X \to E$ such that*

$$\|F(x) - F_\varepsilon(x)\| < \varepsilon, \ \ \textit{for every } x \in X.$$

Theorem 2. [1] *Let (X, d) be a metric space, $(E, \| \cdot \|)$ a Banach space and $F : X \to E$ a proper map. Assume further that for each $\epsilon > 0$, a proper map $F_\epsilon : X \to E$ is given, and the following two conditions are satisfied:*

(a) $\|F_\epsilon(x) - F(x)\| < \epsilon$, *for every $x \in X$,*
(b) for every $\epsilon > 0$ and $u \in E$ in a neighborhood of the origin such that $\|u\| \leq \epsilon$, the equation $F_\epsilon(x) = u$ has exactly one solution x_ϵ.

Then the set $S = F^{-1}(0)$ is an R_δ-set.

3 Main Result

Definition 8. The Wright function $\Psi_\alpha(\theta)$ defined by

$$M_q(\theta) = \sum_{n=1}^{\infty} \frac{(-\theta)^{n-1}}{(n-1)!\Gamma(1-\alpha n)}$$

is such that

$$\int_0^\infty \theta^\delta \Psi_\alpha(\theta) d\theta = \frac{\Gamma(1+\delta)}{\Gamma(1+\alpha\delta)}, \ \text{for } \delta \geq 0.$$

Using definitions 1 and 2, it is very easy to check that the problem (1)–(2) in the equivalent integral equation

$$x(t) = x_0 + h(x) + \frac{1}{\Gamma(q)} \int_0^t (t-s)^{\alpha-1} \\ \times \left[Ax(s) + f(s, x(s)) + \int_0^s K(s-r)x(r)dr \right] ds, \text{ for } t \in [0, b]. \quad (4)$$

provided that the integral in (4) exists. Applying the Laplace transform,
$v(\lambda) = \int_0^\infty e^{-\lambda s} x(s)ds$, and $w(\lambda) = \int_0^\infty e^{-\lambda s} \left(f(s, x(s)) + \int_0^s B(s-r)x(r)dr \right) ds$, $\lambda > 0$,
we can reasoning similarly as in [10, 11] to obtain

$$x(t) = \int_0^\infty \Psi_\alpha(\theta)T(t^\alpha\theta)x_0 d\theta + \alpha \int_0^t \int_0^\infty \theta(t-s)^{\alpha-1}\Psi_\beta(\theta)T((t-s)^\alpha\theta) \\ \times \left[f(s, x(s)) + \int_0^s B(s-r)x(r)dr \right] d\theta ds,$$

Define the operators $\mathscr{K}_\alpha$, $\mathscr{S}_\alpha$

$$\mathscr{K}_\alpha(t) = t^{\alpha-1}P_\alpha(t), \; P_\alpha(t) = \alpha \int_0^\infty \theta\Psi_\alpha(\theta)S((t)^\alpha\theta)d\theta,$$

$$\mathscr{S}_\alpha = \int_0^\infty \Psi_\alpha(\theta)T(t^\alpha\theta)d\theta,$$

The properties of these operators was explored by Zhou [10, 11].
We can also write

$$x(t) = \mathscr{S}_\alpha(t)x_0 + \int_0^t \mathscr{K}_\alpha(t-s) \left[f(s, x(s)) + \int_0^s B(s-r)x(r)dr \right] ds.$$

We need to make the following assumptions.

(H1) $T(t)$ is continuous in the uniform operator topology for $t > 0$, and $\{T(t)\}_{t\geq 0}$f is uniformly bounded, i.e., there exists $M > 1$ such that

$$\sup_{t\in[0,+\infty)} |T(t)| < M.$$

(H2) The map $f : [0, b] \times E \to E$ is satisfies the Carathéodory conditions; that is, $f(., x)$ is measurable for all $x \in E$ and $f(t, .)$ is continuous for almost all $t \in [0, b]$.

(H3) There exists a function $\rho, \overline{\rho} \in L^1(J, R^+)$ and a nondecreasing continuous function $\Omega, \psi : R^+ \to R^+$ such that such that

$$|f(t, x)| \leq \rho(t)\Omega(|x|), \quad \text{for all } t \in [0, b] \text{ and } x \in E.$$

$$|B(t)x| \leq \overline{\rho}(t)\psi(|x|), \quad \text{for all } t \in [0, b] \text{ and } x \in E.$$

(H4) There exists a function $C_f \in C(J, E), C_B \in L^1(J, R^+)$ such that for each nonempty, bounded set $V \subset E$

$$\nu(f(t, V)) \leq C_f(t)\nu(V), \quad \text{for all } t \in [0, b].$$

$$\nu(B(t)V) \leq C_B(t)\nu(V), \quad \text{for all } t \in [0, b].$$

(H5) There are constants $L_h > 0, C_h$ such that

$$\|h(x_1) - h(x_2)\| \leq L_h\|x_1 - x_2\|, \quad \text{for all } x_1, x_2 \in E.$$

$$\nu(h(V) \leq C_h\nu(V), \quad \text{for all } V \subset C([0, b], E).$$

(H6) There exists a constant R satisfying

$$R \geq \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_hR + |h(0)|) + \frac{Mb^\alpha}{\Gamma(\alpha + 1)}\left[\|\rho\|_{L^1}\Omega(R) + \psi(R)\|\overline{\rho}\|_{L^1}\right].$$

we need the following auxiliary lemmas.

Lemma 4. *[10, 11] Under assumption $(H1)$, $P_\alpha(t)$ is continuous in the uniform operator topology for $t > 0$.*

Lemma 5. *[10, 11] Under assumption (H_1), for any fixed $t > 0$, $\{\mathscr{K}_\alpha(t)\}_{t>0}$ and $\{\mathscr{S}_\alpha(t)\}_{t>0}$, are linear operators, and for any $x \in X$*

$$\|\mathscr{K}_\alpha(t)x\| \leq \frac{Mt^{\alpha-1}}{\Gamma(\alpha)}\|x\|, \ \|\mathscr{S}_\alpha(t)x\| \leq \frac{Mt^{\alpha-1}}{\Gamma(\alpha)}\|x\|$$

Lemma 6. *[10, 11] Under assumption (H_1), $\{\mathscr{K}_\alpha(t)\}_{t>0}$ and $\{\mathscr{S}_\alpha(t)\}_{t>0}$, are strongly continuous, which means that, for any $x \in X$ and $0 < t' < t'' \leq b$ we have*

$$\|\mathscr{K}_\alpha(t')x - \mathscr{K}_\alpha(t'')x\| \to 0, \ \|\mathscr{S}_\alpha(t')x - \mathscr{S}_\alpha(t'')x\| \to 0,$$

as $t', t'' \to 0$,

Theorem 3. *Assume that assumptions (H1)–(H5) hold. If*

$$\frac{C_hMb^\alpha}{\Gamma(\alpha)} + \frac{2Mb^\alpha\left(\|C_f\|_\infty + 4M\|C_B\|_{L^1}\right)}{\Gamma(\alpha + 1)} < 1. \tag{5}$$

then the boundary value problem (1)–(2) has at least one mild solution.

We transform the problem (1)–(2) into a fixed point problem. Consider the operator $N : C([0, b], E) \to C([0, b], E)$ defined by:

$$Nx(t) = \mathscr{S}_\alpha(t)x_0 + \int_0^t \mathscr{K}_\alpha(t-s)\left[f(s, x(s)) + \int_0^s B(s-r)x(r)dr\right]ds.$$

Clearly from lemmas 4, 5, 6 the operator N is well defined and the fixed points of N are solutions to (1)–(2). Thus $\mathscr{F}ixN = S(f, x_0 + h(.))$.
Next, we subdivide the proof into several steps as follows Set $D_R = \{x \in C(J, E), \|x\| \leq R\}$.

Step 1. N **maps** D_R **into itself.**
For each $x \in D_R$ we have for each $t \in J$, Using conditions (H_1)–(H_3) and (H_6).

$$\begin{aligned}\|Nx(t)\| &\leq \|\mathscr{S}_\alpha(t)x_0 + \int_0^t \mathscr{K}_\alpha(t-s)\left[f(s, x(s)) + \int_0^s B(s-r)x(r)dr\right]ds\| \\ &\leq \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_h R + |h(0)|) + \frac{M}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} \\ &\quad \left[\|f(s, x(s))\| + \int_0^s \|B(s-r)(x(s))\|dr\right]ds \\ &\leq \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_h R + |h(0)|) + \frac{Mb^\alpha}{\Gamma(\alpha+1)}\left[\|\rho\|_{L^1}\Omega(\|x\|) + \psi(\|x\|)\|\overline{\rho}\|_{L^1}\right] \\ &\leq \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_h R + |h(0)|) + \frac{Mb^\alpha}{\Gamma(\alpha+1)}\left[\|\rho\|_{L^1}\Omega(R) + \psi(R)\|\overline{\rho}\|_{L^1}\right] \leq R.\end{aligned}$$

Step 2. N is continuous.
Let $\{x_n\}$ be a sequence such that $x_n \to x$ in $C([0, b], E)$. The

$$\begin{aligned}&\|N(x_n)(t) - N(x)(t)\| \\ &\leq \|\int_0^t \mathscr{K}_\alpha(t-s)\left[f(s, x_n(s)) - f(s, x(s)) + \int_0^s B(s-r)(x_n(s) - x(s))dr\right]ds\| \\ &\leq \frac{M}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\left[\|f(s, x_n(s)) - f(s, x(s))\| + \|x_n - x\|\|\psi\|_{L^1}\right]ds.\end{aligned}$$

By assumption (H_2) we know that for a.e. $s \in [0, b]$, we have

$$\lim_{n\to+\infty} f(s, x_n(s)) = f(s, x(s)).$$

Hence using dominated convergence theorem, we deduce that

$$\|N(x_n) - N(x)\| \to 0, \quad \text{as } n \to +\infty.$$

Step 3. N maps bounded sets into bounded sets in $C([0, b], E)$. Indeed, it is enough to show that there exists a positive constant ℓ such that for each $x \in B_\eta = \{x \in C([0, b], E) : \|x\| \leq \eta\}$ one has $\|N(x)\| \leq \ell$. Let $x \in B_\eta$. Then for each $t \in [0, b]$, by (H_1), $(H_3)$and (H_5) we have

$$
\begin{aligned}
\|Nx(t)\| &\le \|\mathscr{S}_\alpha(t)x_0 + \int_0^t \mathscr{K}_\alpha(t-s)\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right]ds\| \\
&\le \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_h\eta + |h(0)|) + \frac{M}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1} \\
&\quad \left[\|f(s,x(s))\| + \int_0^s \|B(s-r)(x(s))\|dr\right]ds \\
&\le \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_h\eta + |h(0)|) + \frac{Mb^\alpha}{\Gamma(\alpha+1)}\left[\|\rho\|_{L^1}\Omega(\|x\|) + \psi(\|x\|)\|\overline{\rho}\|_{L^1}\right] \\
&\le \frac{Mb^\alpha}{\Gamma(\alpha)}(x_0 + L_h\eta + |h(0)|) + \frac{Mb^\alpha}{\Gamma(\alpha+1)}\left[\|\rho\|_{L^1}\Omega(\eta) + \psi(\eta)\|\overline{\rho}\|_{L^1}\right] := \ell.
\end{aligned}
$$

Step 4. N maps bounded sets into equicontinuous sets.

We prove $\{Nx,\ x \in B_\eta\}$is equicontinuous
Let $t_1, t_2 \in [0, b]$, $t_1 \le t_2$, let B_η be a bounded set in $C([0, b], E)$ as in Step 2, and let $x \in B_\eta$, we have

$$
\begin{aligned}
&\|N(x)(t_2) - N(x)(t_1)\| \le \\
&\|\mathscr{S}_\alpha(t_2)(x_0 + h(x)) - \mathscr{S}_\alpha(t_1)(x_0 + h(x)) \\
&+ \int_0^{t_2} \mathscr{K}_\alpha(t_2 - s)\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right] \\
&- \mathscr{K}_\alpha(t_1 - s)\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right]ds\| \\
&\le \|\mathscr{S}_\alpha(t_2)(x_0 + h(x)) - \mathscr{S}_\alpha(t_1)(x_0 + h(x))\| \\
&+ \|\int_0^{t_2} \|(\mathscr{K}_\alpha(t_2 - s) - \|\mathscr{K}_\alpha(t_1 - s))\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right]ds \\
&+ \int_{t_1}^{t_2} \mathscr{K}_\alpha(t_1 - s)\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right]ds\| \\
&\le \|\mathscr{S}_\alpha(t_2)(x_0 + h(x)) - \mathscr{S}_\alpha(t_1)(x_0 + h(x))\| \\
&+ \|\int_0^{t_2} \|(\mathscr{K}_\alpha(t_2 - s) - \|\mathscr{K}_\alpha(t_1 - s))\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right]ds\| \\
&+ \|\int_{t_1}^{t_2} \mathscr{K}_\alpha(t_1 - s)\left[f(s,x(s)) + \int_0^s B(s-r)x(r)dr\right]ds\|.
\end{aligned}
$$

From the fact that $\mathscr{S}_\alpha(t)$, $\mathscr{K}_\alpha(t)$ are uniformly continuous on J, we deduce then $\{Nx,\ x \in B_\eta\}$ is equi-continuous. Now let V be a subset of D_R such that $V \subset conv(T(V) \cup \{0\})$. V is bounded and equicontinuous, and therefore the function $t \to v(t) = \nu(V(t))$ is continuous and bounded on J.

From properties of the measure of noncompactness ν, we have for each $t \in J$,

$$
\begin{aligned}
v(t) &\le \nu(N(V)(t)) \cup \{0\}) \\
&\le \nu(N(V)(t)).
\end{aligned}
$$

First we will estimate $\nu(N(V)(t))$.

Using Lemma 1, (H_4), (H_5) and the properties of the measure of noncompactness ν one has,

$$\begin{aligned}\nu(NV(t)) &\le \nu\left\{\mathscr{S}_\alpha(t)(x_0+h(x))+\int_0^t \mathscr{K}_\alpha(t-s)\right.\\ &\qquad \left.\left[f(s,x(s))+\int_0^s B(s-r)x(r)dr\right]ds,\ x(t)\in V(t)\right\}\\ &\le \nu\left\{\mathscr{S}_\alpha(t)(x_0+h(x))\ x(t)\in V(t)\right\}\\ &+\nu\left\{\int_0^t \mathscr{K}_\alpha(t-s)\left[f(s,x(s))+\int_0^s B(s-r)x(r)dr\right]ds,\ x(s)\in V(t)\right\}\\ &\le \frac{C_hMb^\alpha}{\Gamma(\alpha)}\nu(V(t))+\frac{M}{\Gamma(\alpha)}\int_0^t (t-s)^{\alpha-1}\left[C_f(t)\nu(V(t))+\|C_B\|_{L^1}\nu(V(t)\right]ds\\ &\le \left[\frac{C_hMb^\alpha}{\Gamma(\alpha)}+\frac{Mb^\alpha\left(\|C_f\|_\infty+\|C_B\|_{L^1}\right)}{\Gamma(\alpha+1)}\right]\nu(V(t))\end{aligned}$$

It follows then that

$$v(t)\le \|v\|_\infty\left[\frac{C_hMb^\alpha}{\Gamma(\alpha)}+\frac{Mb^\alpha\left(\|C_f\|_\infty+\|C_B\|_{L^1}\right)}{\Gamma(\alpha+1)}\right],$$

which means that

$$\|v\|_\infty\left(1-\frac{C_hMb^\alpha}{\Gamma(\alpha)}+\frac{Mb^\alpha\left(\|C_f\|_\infty+\|C_B\|_{L^1}\right)}{\Gamma(\alpha+1)}\right)\le 0.$$

By (5) it follows that $\|v\|_\infty=0$; that is, $v(t)=0$ for each $t\in J$, and then $V(t)$ is relatively compact in E. In view of the Ascoli-Arzela theorem, V is relatively compact in D_R. Applying now Theorem 1, we conclude that N has a fixed point which is a solution of the problem (1)–(2).

4 Compactness of Solution Set

Now we show that the set

$$S=\{x\in C([0,T],E): X \text{ is a solution of } (1)-(2) \text{ is compact}\}.$$

Let $(x_n)_{n\in N}$ be a sequence in S. We put $\Lambda=\{x_n : n\in N\}\subseteq C([0,b],E)$. Then from earlier parts of the proof of this theorem, we conclude that Λ is bounded and equicontinuous, in another hand it easily to prove that the set

$$\{x_n(t), n\in N,\ x_n\in\Lambda\},$$

for a fixed point $t \in J$ is relatively compact. Then from the Ascoli-Arzela theorem, we can conclude that Λ is compact, it follows then $\{x_n\}_{n \in N}$ has a subsequence $(x_{n_m})_{n_m \in N}$ converges to x with $(x_{n_m})_{n_m \in N} \subset S = \{x \in C([0, b], E) : x$ is a solution of $(1) - (2)\}$.
Let

$$z(t) = \mathscr{S}_\alpha(t)(x_0 + h(x)) + \int_0^t \mathscr{K}_\alpha(t - s) \left[f(s, x(s)) + \int_0^s B(s - r)x(r)dr \right] ds$$

and

$$\begin{aligned} |x_{n_m}(t) - z(t)| \leq & |\mathscr{S}_\alpha(t)(x_0 + h(x_m)) - \mathscr{S}_\alpha(t)(x_0 + h(x))| \\ & + \int_0^t \mathscr{K}_\alpha(t - s) \left[|f(s, x_m(s)) - f(s, x(s))| \right. \\ & \left. + \int_0^s \|B(s - r)\| |x_m(r) - x(s)| dr \right] ds. \end{aligned}$$

As $n_m \to +\infty$, $x_{n_m}(t) \to z(t)$, and then

$$x(t) = \mathscr{S}_\alpha(t)(x_0 + h(x)) + \int_0^t \mathscr{K}_\alpha(t - s) \left[f(s, x(s)) + \int_0^s B(s - r)x(r)dr \right] ds.$$

We conclude that $\{x_n \mid n \in N\}$ has subsequence converging to x in S.

Hence $S(f, x_0 + h(x))$ is compact.

Define

$$\tilde{f}(t, x(t)) = \begin{cases} f(t, x(t)), |x(t)| \leq M, \\ f(t, \frac{Mx(t)}{|x(t)|}), |x(t)| \geq M, \end{cases}$$

$$\tilde{B}(t, x(t)) = \begin{cases} B(t, x(t)), |x(t)| \leq M, \\ B(t, \frac{Mx(t)}{|x(t)|}), |x(t)| \geq M, \end{cases}$$

$$\tilde{h}(x) = \begin{cases} h(x), \|xt\| \leq M, \\ h(\frac{Mx}{\|x\|}), \|x\| \geq M, \end{cases}$$

Since f, B, h are continuous, the function $\widetilde{f}, \widetilde{B}, \widetilde{h}$ are continuous and bounded by (H_4), (H_5) so there exists $M_f > 0$, $M_B > 0$ and $M_h > 0$ such that

$$\widetilde{f}(t, u)| \leq M_f, \text{ for } a.e.\ t \text{ and all } u \in E. \tag{6}$$

$$|\widetilde{B}(t)u| \leq M_B, \text{ for } a.e.\ t \text{ and all } u \in E. \tag{7}$$

$$|\widetilde{h}x| \leq M_h, \text{ for } a.e.\ t \text{ and all } x \in C(J, E). \tag{8}$$

We consider the following modified problem,

$$\begin{cases} {}^cD^{\alpha}x(t) = Ax(t) + \int_0^t \widetilde{B}(t-s)x(s)ds + \widetilde{f}(t, x(t)),\ t \in J = [0, b], \quad 0 < \alpha \leq 1, \\ x(0) = x_0 + \widetilde{h}(x). \end{cases}$$

We can easily prove that $S(f, B, x_0 + h(x)) = S(\tilde{f}, \widetilde{B}x_0 + h(x)) = Fix\tilde{N}$, where

$$\tilde{N} : C([0, b], E) \longrightarrow C([0, b], E)$$

is defined by

$$\tilde{N}(x)(t) = \mathscr{S}_\alpha(t)(x_0 + \widetilde{h}(x)) + \int_0^t \mathscr{K}_\alpha(t-s)\left[\widetilde{f}(s, x(s)) + \int_0^s \widetilde{B}(s-r)x(r)dr\right]ds.$$

We can estimate

$$\begin{aligned} \tilde{N}(x)(t) &\leq \frac{Mt^\alpha}{\Gamma(\alpha)}|x_0 + \widetilde{h}(x)| + \int_0^t \mathscr{K}_\alpha(t-s)\left[\widetilde{f}(s, x(s)) + \int_0^s \widetilde{B}(s-r)x(r)dr\right]ds \\ &\leq \frac{Mt^\alpha}{\Gamma(\alpha)}|x_0 + \widetilde{h}(x)| + \frac{Mt^\alpha}{\Gamma(\alpha+1)}[M_f + bM_B] \\ &\leq \frac{Mt^\alpha}{\Gamma(\alpha)}(|x_0| + M_h) + \frac{Mb^\alpha}{\Gamma(\alpha+1)}[M_f + bM_B]. \end{aligned}$$

Finally we have

$$\|\tilde{N}(x)\| \leq M^*,$$

then $\tilde{N}$ is uniformly bounded, as in steps 3 to 4 we can prove that

$$\tilde{N} : C([0, b], E) \longrightarrow C([0, b], E),$$

is compact which allows us to define the compact perturbation of the identity
$\widetilde{G}(x) = x - \tilde{N}(x)$ which is a proper map.

From the compactness of $\widetilde{N}$, we can easily prove that all conditions of Theorem (2) are satisfied. Therefore the solution set $S(\tilde{f}, c_0) = \tilde{G}^{-1}(0)$ is an R_δ-set so $S(f, c_0)$ is an R_δ-set.

5 Conclusion

In this paper we have given a result concerning the existence of Mild solution of a class of fractional partial differential integro-differential equation with nonlocal conditions using measure of non-compactness combined with Mönch fixed point theorem, we also extend the classical Kneser's theorem and Aronszajn type result for this class of equations by showing that the set of all solutions is compact and R_δ-set.

Acknowledgements The author would like to express his warmest thanks to all members of $ICDDEA19$ International Conference on Differential, Difference Equations and Applications for his/her valuable comments, helps and suggestions.

References

1. Djebali, S., Górniewicz, L., Ouahab, A.: Solutions Sets for Differential Equations and Inclusions. De Gruyter, Berlin (2013)
2. Cannarsan, P., Sforza, D.: Global solutions of abstract semilinear parabolic equations with memory terms. Nonlinear Differ. Equ. Appl. **10**, 399–430 (2003)
3. Ezzinbi, K., Ghnimi, S., Taoudi, M.A.: Existence and regularity of solutions for neutral partial functional integrodifferential equations with infinite delay. Nonlinear Anal. Hybrid Syst. **4**, 54–64 (2010)
4. Ezzinbi, K., Fu, X.: Existence and regularity of solutions for some neutral partial differential equations with nonlocal conditions. Nonlinear Anal. **57**, 1029–1041 (2004)
5. Ezzinbi, K., Fu, X., Hilal, K.: Existence and regularity in the a-norm for some neutral partial differential equations with nonlocal conditions. Nonlinear Anal. **67**, 1613–1622 (2007)
6. Ezzinbi, K., Taoudi, M.A.: Sadovskii-Krasnosel'skii type fixed point theorems in Banach spaces with application to evolution equations. J. Appl. Math. Comput. **49**, 243–260 (2015)
7. Fu, X., Ezzinbi, K.: Existence of solutions for neutral functional differential evolution equations with nonlocal conditions. Nonlinear Anal. **54**, 215–227 (2003)
8. Kamenskii, M., Obukhovskii, V., Zecca, P.: Condensing Multivalued Maps and Semilinear Differential Inclusions in Banach Spaces. De Gruyter, Berlin (2001)
9. Mönch, H.: Boundary value problems for nonlinear ordinary differential equations of second order in Banach spaces. Nonlinear Anal. **4**, 985–999 (1980)
10. Zhou, Y.: Basic Theory of Fractional Differential Equations. World Scientific, Singapore (2014)
11. Zhou, Y.: Fractional Evolution Equations and Inclusions, Analysis and Control. Elsevier, Amsterdam (2015)

Study of the Equivalent Impedance of a Resonator Array Using Difference Equations

Helena Albuquerque and José Alberto

Abstract In this paper we present a mathematical model, based on the theory of difference equations, to study the wireless power transfer using resonator arrays. This approach is relevant for a better understanding of the system because it gives us a closed form for the equivalent impedance of the array of resonators that allows us to make predictions about the behavior of the system. We prove that the complete solution of the problem depends only on the initial conditions and possible perturbations introduced. Using the computer software MATLAB we present in the last section of this paper some illustrations of our results.

Keywords Wireless power transfer · Difference equations · Resonator array

1 Introduction

Recently, wireless power transfer (WPT) has received considerable attention in the scientific community due to a significant number of applications, ranging from charging electrical vehicles [1] to powering small electronic devices [9, 11]. These systems are able to transfer power without needing electrical contact, which is particularly useful in harsh environments (dust, dirt or water). In order to transfer power over longer distances, arrays of magnetically coupled resonators are used [2, 3, 5, 7, 8, 10, 12, 13], placed in a line with a receiver that slides along the line. The electrical

Supported by the Centre for Mathematics of the University of Coimbra—UID/MAT/00324/2019, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

H. Albuquerque (✉)
CMUC, Departamento de Matemática, Universidade de Coimbra,
Apartado 3008, 3001-454 Coimbra, Portugal
e-mail: lena@mat.uc.pt

J. Alberto
Institute of Systems and Robotics, Department of Electrical and Computer Engineering,
University of Coimbra, 3030-290 Coimbra, Portugal
e-mail: jose.alberto@isr.uc.pt

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_25

characterization of the resonator array is needed to design an efficient WPT system. More specifically, the knowledge of the value of the resonator array equivalent impedance allows the power supplied by the source to be predicted.

In this paper we develop the mathematical approach described in [3, 4, 14] in order to study the continued fraction that represents the equivalent impedance of the array of resonators as a recursive sequence, which general term can be determined through the resolution of finite difference equations with constant coefficients. Although we briefly approached the behavior of the system for large numbers, we will deepen this study in this work, presenting a description of the monotonicity and of the speed of the convergence of the modeling sequence. To illustrate the mathematical approach developed in this paper, a few examples made with the software MATLAB are presented. Moreover, the approach developed in this work gives a consistent theoretical basis that can be used as a powerful tool for designing a WPT system composed of an array of resonators with chosen properties and behaviour.

2 Description of the Circuit

In this work we will consider the same circuit described in [2–4]: an array of resonators (cells) disposed in a line. Each two adjacent resonators are spaced by the same distance between them and are magnetically coupled with a mutual inductance M, while the coupling between nonadjacent resonators is neglected. Each cell can be described as an R-L-C series circuit, as seen in Fig. 1 [4]. In the circuit, R represents the intrinsic resistance of the resonator cell, L its self-inductance and C the additional capacitance needed to tune it to the resonant frequency $\omega_0 = 2\pi f_0 = 1/\sqrt{LC}$. The impedance of each cell is then given by: $\hat{Z} = R + j\omega L + 1/(j\omega C)$, which, at the resonant angular frequency ω_0, becomes equal to its intrinsic resistance R. A voltage source $\hat{V}_s$ feeds the first cell of the array and a termination impedance $\hat{Z}_T$ is be connected to the last cell of the array.

When there is a receiver above the line, as shown in Fig. 1(a), the receiver absorbs part of the power arriving from the source at the lth cell under it and this fact represents a perturbation in the system. This perturbation caused by the magnetic coupling between the receiver and the lth cell of the resonator array adds an impedance $\hat{Z}_d$ to that cell (lth cell). So the equivalent circuit can be represented as in Fig. 1(b). We assume that $\hat{Z}_d = R_d$ is real, when working at the resonant frequency, as we consider the receiver has the same resonant frequency as the cells of the array.

The resonator array circuit can be simplified by using an equivalent impedance $\hat{Z}_{eq}$, that represents the impedance of all the resonators after the supplied one and the receivers (if any), as depicted in Fig. 2. The equivalent impedance $\hat{Z}_{eq}$ has the

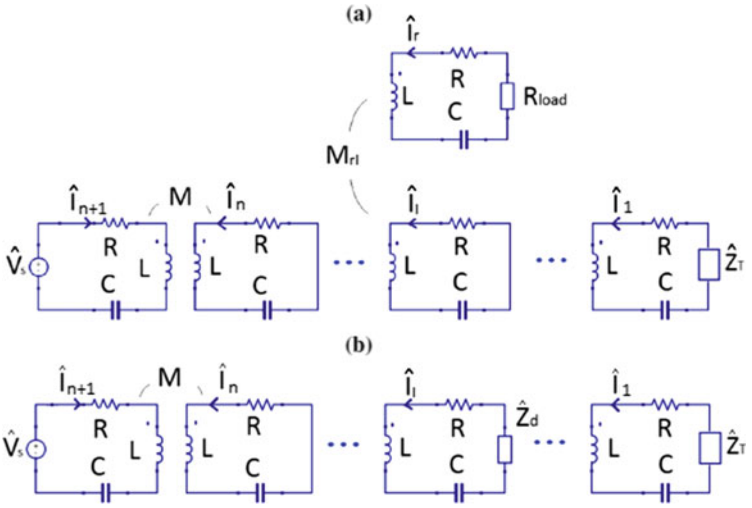


Fig. 1 Circuit of (a) an array composed of $n+1$ cells with a receiver above the lth cell and (b) the equivalent circuit with the impedance $\hat{Z}_d$ representing the receiver inserted in the lth cell

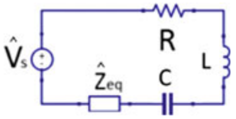


Fig. 2 Equivalent circuit for the array in Fig. 1 with an equivalent impedance $\hat{Z}_{eq}$ representing the resonator array and the receiver (if any), excluding the resonator connected to the voltage source

following expression, if the receiver is placed above the last cell of the resonator array ($\hat{Z}'_T = \hat{Z}_T + \hat{Z}_d$) or if there is no receiver ($\hat{Z}'_T = \hat{Z}_T$) [2–4]:

$$\hat{Z}_{eq} = \cfrac{(\omega M)^2}{\hat{Z} + \cfrac{(\omega M)^2}{\cdots + \cfrac{(\omega M)^2}{\hat{Z} + \cfrac{(\omega M)^2}{\hat{Z} + \hat{Z}'_T}}}}. \tag{1}$$

In case the receiver coil is placed above any other cell of the resonator array, introducing the impedance $\hat{Z}_d$, $\hat{Z}_{eq}$ becomes:

$$\hat{Z}_{eq} = \cfrac{(\omega M)^2}{\hat{Z} + \cfrac{(\omega M)^2}{\cdots + \cfrac{(\omega M)^2}{\hat{Z}_d + \hat{Z} + \cfrac{(\omega M)^2}{\cdots + \cfrac{(\omega M)^2}{\hat{Z} + \hat{Z}_T}}}}}. \tag{2}$$

3 Mathematical Analysis of the Continued Fraction

In the following section we will describe the mathematical approach described in [3, 4], by developing the study of the system for large natural values, in particular regarding monotonicity and the speed of convergence of the sequence. We prove in this section that, although the behavior of the system for very large values does not depend on the initial conditions and on the introduced perturbations, the opposite happens with the monotonicity (in the real case) and the convergence speed of the modeling sequence. Those depend on the initial conditions of the problem but, as it will be explained later in this section, they can be controlled by predictable limits.

3.1 Fraction Without a Perturbation

We can rewrite the continued fraction (1) for any number $n + 1$ of resonators and $n \geq 0$ using generic letters a, $b \in \mathbb{C}^*$:

$$x_n = \frac{a}{b+} \frac{a}{b+} \cdots \frac{a}{b}. \tag{3}$$

Then, (3) is the nth term of the following recursive sequence (with $k \geq 1$):

$$x_k = \frac{a}{b + x_{k-1}} \tag{4}$$

with

$$x_0 = \frac{p_0}{q_0}. \tag{5}$$

Knowing that the $(n + 1)$ cells of the array are labelled from 1 to $n + 1$, with 1 being the cell connected to the termination impedance and $n + 1$ is the cell connected to the voltage source (Fig. 1), the term x_0 corresponds to the termination impedance of the array ($\hat{Z}'_T$ in (1) or $\hat{Z}_T$ in (2)). Thus, noting by

$$x_k = \frac{p_k}{q_k}, \tag{6}$$

we can verify by induction that $\{p_n\}$ and $\{q_n\}$ are sequences defined by the following recurrence relations:

$$\begin{array}{l} p_n = bp_{n-1} + ap_{n-2} \\ q_n = bq_{n-1} + aq_{n-2} \end{array} \text{for } \ n \geq 2, \tag{7}$$

being p_0 and q_0 fixed and p_1 and q_1 given by

$$p_1 = aq_0; \ q_1 = bq_0 + p_0 \, . \tag{8}$$

Thus, the sequences $\{p_n\}$ and $\{q_n\}$ are defined by linear homogeneous second order difference equations with constant coefficients which can be solved directly. In fact, the equation in (7) given by $p_n - bp_{n-1} - ap_{n-2} = 0$ is a linear homogeneous second order difference equation with constant coefficients. Therefore, from [6] its solution is given by $p_n = m_1\lambda_1^n + m_2\lambda_2^n$ supposing that λ_1 and λ_2 are distinct solutions of the equation $\lambda^2 - b\lambda - a = 0$ (which is the case that applies to our practical application):

$$\lambda^2 - b\lambda - a = 0 \Leftrightarrow \lambda = \frac{b \pm \sqrt{b^2 + 4a}}{2} \tag{9}$$

where m_1 and m_2 are constants that should be determined using the initial conditions. The same considerations can be done for $\{q_n\}$, considering $q_n = m_3\lambda_1^n + m_4\lambda_2^n$.

Then we can determine the general term of the sequence $\{x_n\} = \{p_n/q_n\}$:

$$x_n = \frac{a_1\left(\frac{b-\sqrt{b^2+4a}}{2}\right)^n + a_2\left(\frac{b+\sqrt{b^2+4a}}{2}\right)^n}{b_1\left(\frac{b-\sqrt{b^2+4a}}{2}\right)^n + b_2\left(\frac{b+\sqrt{b^2+4a}}{2}\right)^n} \tag{10}$$

in which a_1, a_2, b_1 and b_2 are constants determined by the initial conditions. For simplicity, setting $p_0 = x_0$ and q_0=1, a_1, a_2, b_1 and b_2 can be obtained by solving the following system:

$$\begin{array}{l} a_1 + a_2 = x_0 \\ b_1 + b_2 = 1 \\ \frac{a_1}{2}\left(b - \sqrt{b^2 + 4a}\right) + \frac{a_2}{2}\left(b + \sqrt{b^2 + 4a}\right) = a \\ \frac{b_1}{2}\left(b - \sqrt{b^2 + 4a}\right) + \frac{b_2}{2}\left(b + \sqrt{b^2 + 4a}\right) = b + x_0 \end{array} \tag{11}$$

3.2 Fraction with a Perturbation (l^{th} Term)

The expression (2), is the equivalent impedance of a multiple resonator system for a receiver placed above the lth cell of the receiver line and can be rewritten using the following fraction:

$$x_n = x_n = \frac{a}{b+}\frac{a}{b+}\cdots\frac{a}{b'+}\cdots\frac{a}{b} \quad \text{with } a,\ b,\ b' \in \mathbb{C}^*. \tag{12}$$

To solve this fraction and determine its value, we split the fraction in two fractions (one with l and the other with $n-l$ terms), determine the value of x_k for the $(l-1)$th term, then determine the lth value x_l using the perturbation b' and, finally, using x_l as an initial value, we determine the value of the fraction, with the last $n-l$ values.

We start by calculating the term of $(l-1)$th order, using (10):

$$x_{l-1} = \frac{a_1\left(b-\sqrt{b^2+4a}\right)^{l-1} + a_2\left(b+\sqrt{b^2+4a}\right)^{l-1}}{b_1\left(b-\sqrt{b^2+4a}\right)^{l-1} + b_2\left(b+\sqrt{b^2+4a}\right)^{l-1}} \tag{13}$$

in which a_1, a_2, b_1 and b_2 are determined by the initial conditions x_0 and x_1 as done before, with (11). After, we have to determine the value of

$$x_l = \frac{a}{b' + x_{l-1}} = y_0 \tag{14}$$

and

$$\frac{a}{b + y_0} = y_1. \tag{15}$$

Then, (14) and (15) are the initial conditions used to determine the value of the fraction $y_{n-l} = x_n$:

$$y_{n-l} = x_n = \frac{c_1\left(b-\sqrt{b^2+4a}\right)^{n-l} + c_2\left(b+\sqrt{b^2+4a}\right)^{n-l}}{d_1\left(b-\sqrt{b^2+4a}\right)^{n-l} + d_2\left(b+\sqrt{b^2+4a}\right)^{n-l}}. \tag{16}$$

The constants c_1, c_2, d_1 and d_2 (shown in appendix) are determined by the initial conditions y_0 and y_1, using (11), where, for simplicity is assumed that $p_0 = y_0$ and $q_0 = 1$.

In conclusion, (16) represents the value of the fraction (2) for $n+1$ resonators with the perturbation in the lth term, (receiver facing the lth resonator, in which $l=1$ resonator connected to the termination impedance and $l=n$ in the resonator after the one connected to the voltage source).

3.3 Study of the Convergence of the Continued Fraction

Following the study of the fraction, we can study its behaviour for a large number of terms, in other words its value for $n \to \infty$.

Supposing that $z^2 - bz - a = 0$ has distinct roots z_1 and z_2, with $|z_1| > |z_2|$, we have

$$\left|\frac{z_2}{z_1}\right| < 1, \text{ so } \lim_{n\to\infty}\left(\frac{z_2}{z_1}\right)^n = 0. \tag{17}$$

Now we can demonstrate that x_n tends to the quotient of the coefficients of z_1^n. For example, if $\left|b - \sqrt{b^2+4a}\right| < \left|b + \sqrt{b^2+4a}\right|$, i.e. $z_1 = b + \sqrt{b^2+4a}$ and $z_2 = b - \sqrt{b^2+4a}$ from (10) we can conclude that

$$x_n = \frac{a_2 z_1^n + a_1 z_2^n}{b_2 z_1^n + b_1 z_2^n} \tag{18}$$

thus

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \frac{a_2 + a_1\left(\frac{z_2}{z_1}\right)^n}{b_2 + b_1\left(\frac{z_2}{z_1}\right)^n} = \frac{a_2}{b_2}. \tag{19}$$

Calculating the constants a_2 and b_2, it is interesting to note that the value of the limit (19) is always equal to $\frac{1}{2}\left(\sqrt{4a+b^2} - b\right)$ and that does not depend on the initial conditions. This means that, for fixed values of a, $b \in \mathbb{C}$, the value of the fraction is always the same when the number of terms is infinite. So if we set the value $x_0 = \frac{1}{2}\left(\sqrt{4a+b^2} - b\right)$ it is easy to prove that $\{x_n\}$ becomes a constant sequence. Furthermore, for a finite number of perturbations, the behaviour of the fraction at infinity remains the same.

3.4 Study of the Monotonicity of the Sequence

Following the analysis of the convergence, for $a, b, b', p_0, q_0 \in \mathbb{R}$, with $a, b > 0$ and $x_0 \geq 0$, we study the monotonicity of the sequence, i.e., whether the sequence that represents the continued fraction increases or decreases with the increase of the number terms.

From (18) we can write that

$$x_n = \frac{a_1}{b_1} + \frac{a_2 b_1 - a_1 b_2}{b_1\left(b_1\left(\frac{z_2}{z_1}\right)^n + b_2\right)} \tag{20}$$

We now study the monotonicity of $\{b_1 w_n + b_2\}$, with $w_n = \left(\frac{z_2}{z_1}\right)^n$. The alternate behaviour of $\{w_n\}$ with positive and negative values, being its sub-sequences $\{w_{2n}\}$ and $\{w_{2n+1}\}$ with different monotonicities (one is increasing, the other is decreasing) and both convergent to zero, yields an analogous behaviour of the sequence $\{b_1 w_n + b_2\}$. The sign of the sequence depends on the sign of b_1 that is positive if $x_0 < \frac{a_2}{b_2}$ and negative otherwise, as b_2 is always positive.

Thus, in case $x_0 > \frac{a_2}{b_2}$, the subsequences $\{b_1 w_{2n} + b_2\}$ and $\{b_1 w_{2n+1} + b_2\}$ are decreasing and increasing with values higher and lower than b_2, respectively, and both converging to b_2. It can be seen the decrease of an even term to its consecutive odd term and an increase of an odd term to its consecutive even term. The opposite occurs if $x_0 < \frac{a_2}{b_2}$. Moreover, $\{b_1 w_n + b_2\}$ is a sequence with positive real terms, as $|b_1| < |b_2|$. Then, considering that $(a_2 b_1 - a_1 b_2)$ is negative if $x_0 > \frac{a_2}{b_2}$ and positive otherwise, we have that $\frac{a_2 b_1 - a_1 b_2}{b_1}$ is always positive.

Regarding the monotonicity of the sequence we can conclude that is not monotonic and the terms tend to the limit alternatively from lower and higher values, as it converges. Moreover, all the terms of the sequence are bounded by the first two terms, x_0 and x_1.

3.5 Study of the Speed of Convergence

After the study of the monotonicity, we study the speed of convergence by finding the order of the term in which the absolute value of the difference between the value of the fraction and the limit of the fraction is smaller than a given value ε.

Without any lack of generality, we suppose that $\left|b - \sqrt{b^2 + 4a}\right| < \left|b + \sqrt{b^2 + 4a}\right|$. Using (18) and the initial conditions described above, we have that $(a_1 b_2 - a_2 b_1) = \frac{2}{z_1 - z_2}(p_0 q_1 - p_1 q_0)$ and that $(a_1 b_2 - a_2 b_1)$ depends on the initial conditions. Therefore, letting

$$\delta_n = \left|x_n - \lim_{n\to\infty} x_n\right| = \left|\frac{a_2 z_1^n + a_1 z_2^n}{b_2 z_1^n + b_1 z_2^n} - \frac{a_2}{b_2}\right| \tag{21}$$

for any $\varepsilon > 0$ there is an integer number N such that for $n > N$, we have $\delta_n < \varepsilon$. Knowing that $w_n = \left(\frac{z_2}{z_1}\right)^n$ is a sequence that tends to zero due to $\left|\frac{z_2}{z_1}\right| < 1$, we have

$$\delta_n = \left|\frac{a_2 + a_1 w_n}{b_2 + b_1 w_n} - \frac{a_2}{b_2}\right| = \left|\frac{(a_1 b_2 - a_2 b_1)\, w_n}{b_2\,(b_1 w_n + b_2)}\right| = \frac{|a_1 b_2 - a_2 b_1|\,|w_n|}{|b_2|\,|b_1 w_n + b_2|} \tag{22}$$

Assuming that Q is a non-zero lower bound of the convergent sequence $|b_1 w_n + b_2|$, we have that $|b_1 w_n + b_2| > Q$, so

$$\delta_n = \frac{|a_1 b_2 - a_2 b_1|\,|w_n|}{|b_2|\,Q} < \varepsilon \tag{23}$$

is equivalent to

$$N > \log_{\left|\frac{z_2}{z_1}\right|}\left(\frac{|b_2|\,Q}{|a_1 b_2 - a_2 b_1|}\varepsilon\right) \tag{24}$$

In conclusion, for any $\varepsilon > 0$, we can set an order N equal to the largest integer contained in $\log_{\left|\frac{z_2}{z_1}\right|}\left(\frac{|b_2|Q}{|a_1 b_2 - a_2 b_1|}\varepsilon\right)$ such that for $n > N$, x_n is in the circle centered in $-\frac{z_2}{2}$ with a radius equal to ε. For the particular case where the constants are real, we can set $Q = |b_1 w_0 + b_2|$ if $x_0 > \frac{a_2}{b_2}$, or $Q = |b_1 w_1 + b_2|$ otherwise.

4 Numerical Analysis of the Value and Characteristics of the Continued fraction

In this section, the theoretical results described in the previous section are applied to a WPT system composed of an array of resonators and the expressions (1) and (2) are found. Examples of calculations are carried out with the software MATLAB and some numerical results are presented and discussed, using the values determined experimentally in [4]: $L = 12.6\,\mu\text{H}$, $C = 93.1\,\text{nF}$, $R = 0.11\,\Omega$, $M = -1.55\,\mu\text{H}$ and $f_0 = 147\,\text{kHz}$.

4.1 Numerical Analysis of the Value of the Equivalent Impedance

To obtain the expression of the equivalent impedance $\hat{Z}_{eq}$, we write the generic values of the fractions (10) and (16) in terms of the characteristics of the WPT system, as done in [4].: $a = (\omega M)^2, b = \hat{Z}, x_0 = \hat{Z}'_T = \hat{Z}_T + \hat{Z}_d$ (which is reduced to $\hat{Z}_T$ when the receiver is not above the last cell), $b' = \hat{Z}_d + \hat{Z}$.

4.1.1 Value of the Equivalent Impedance Without a Receiver over the Resonator line or with the Receiver on the Last Cell of the Resonator Line

$$\hat{Z}_{eq} = \frac{f^n(2(\omega M)^2 - g\hat{Z}'_T) + g^n(f\hat{Z}'_T - 2(\omega M)^2)}{f^n(f + 2\hat{Z}'_T) - g^n(g + 2\hat{Z}'_T)} \tag{25}$$

where $f = \hat{Z} - \sqrt{\hat{Z}^2 + 4\,(\omega M)^2}$ and $g = \hat{Z} + \sqrt{\hat{Z}^2 + 4\,(\omega M)^2}$.

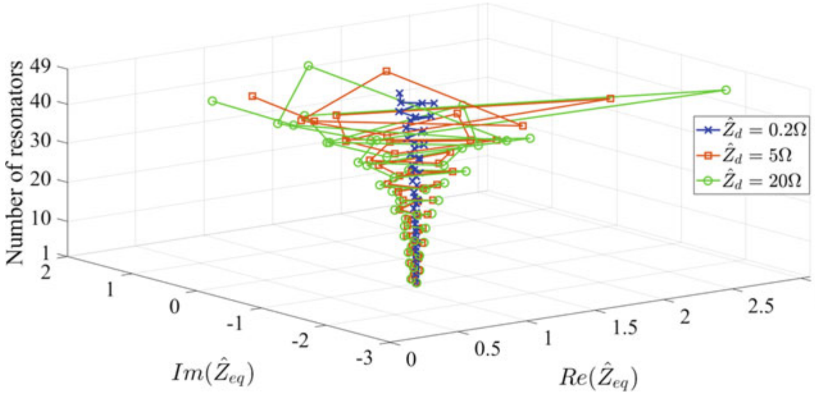


Fig. 3 Value of $\hat{Z}_{eq}$ (real and imaginary parts) for different receiver positions and for different values of $\hat{Z}_d$, for $f = 160\,\text{kHz}$ and for $\hat{Z}_T = 1.5\,\Omega$. The position of the receiver is 1 when is over the cell connected to the termination impedance and 49 when over the cell after the one connected to the voltage source

4.1.2 Value of the Equivalent Impedance with a Receiver over the Resonator Line at Any Position

$$\hat{Z}_{eq} = \frac{(\omega M)^2 \left(e_1 f^n g^{2l} + e_2 f^{2l} g^n - f^l g^l \left(e_3 f^n + e_4 g^n\right)\right)}{f^n g^l \left(e_5 f^l + e_6 g^l\right) + f^l g^n \left(e_7 f^l + e_8 g^l\right)} \tag{26}$$

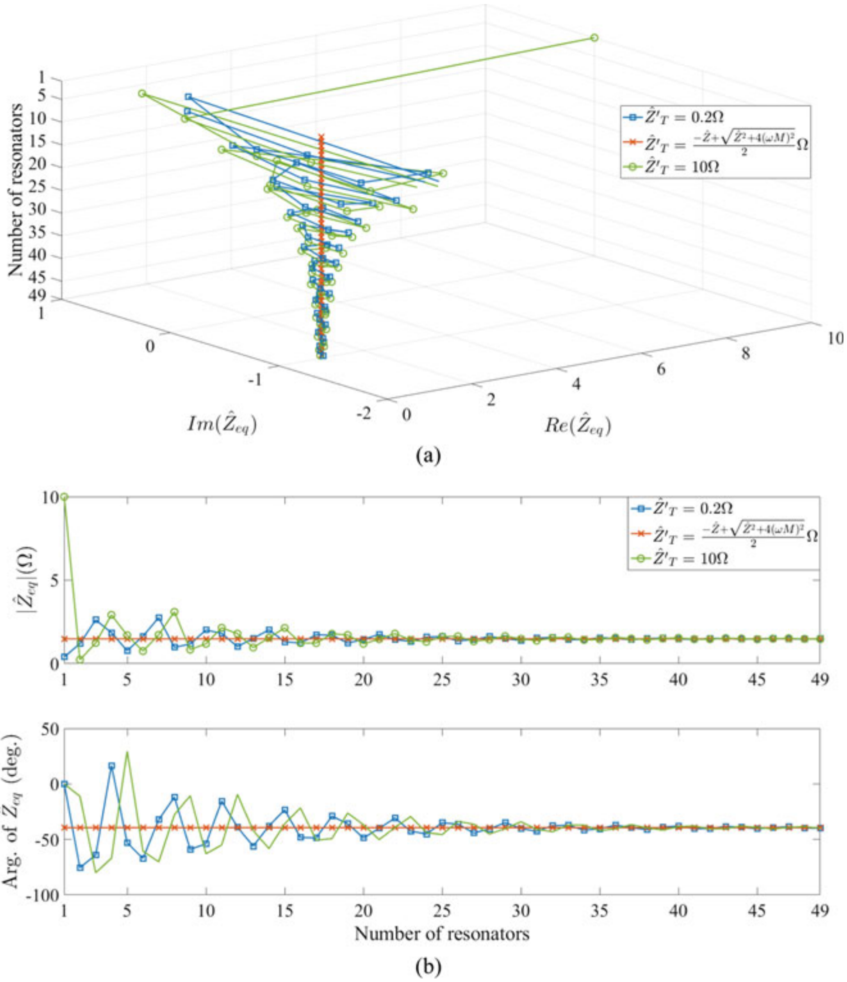
where the constants e_1, e_2, e_3, e_4, e_5, e_6, e_7, and e_8 are described in the Appendix.

In Fig. 3 the equivalent impedance $\hat{Z}_{eq}$ is calculated versus the position of the receiver and for different values of the receiver impedance $\hat{Z}_d$, considering the case in which we are working at a frequency ($f = 160$ kHz) different than the resonant one ($\omega \neq \omega_0$), meaning that $\hat{Z}$ has an imaginary value. We can notice that for a line of 50 resonators, the equivalent impedance is affected more significantly as the receiver gets closer to the cell after the one connected to the source. Moreover, the effect increases with the value of $\hat{Z}_d$.

4.2 Numerical Study of the Convergence of the Fraction

For the constants a_2 and b_2 given in the Appendix, for an infinite number of resonators the continued fraction (25) converges to the following value:

$$\lim_{n\to\infty} \hat{Z}_{eq} = \frac{1}{2}\left(-\hat{Z} + \sqrt{\hat{Z}^2 + 4\,(\omega M)^2}\right). \tag{27}$$



(a)

(b)

Fig. 4 Value of the equivalent impedance $\hat{Z}_{eq}$ with (a) its real and imaginary parts and (b) its magnitude and angle, for $f = 160\,\text{kHz}$, different values of $\hat{Z}'_T$ and different numbers of resonators

We can see that (27) does not depend on the initial conditions, i.e., the impedance $\hat{Z}'_T$. The limit depends only on the electrical characteristics of the cells, the mutual inductance M and the angular frequency ω. Using the expression (1), for a frequency $f = 160\,\text{kHz}$ different than the resonant frequency, we can obtain and plot the equivalent impedance $\hat{Z}_{eq}$ for different numbers of resonators $n + 1$ and for different values of $\hat{Z}'_T$ (Fig. 4).

As we increase the length of the resonator line, the equivalent impedance converges to the same value, even for different values of the termination impedance $\hat{Z}'_T$. Moreover, setting the impedance $\hat{Z}'_T$ equal to (27), the value of the equivalent

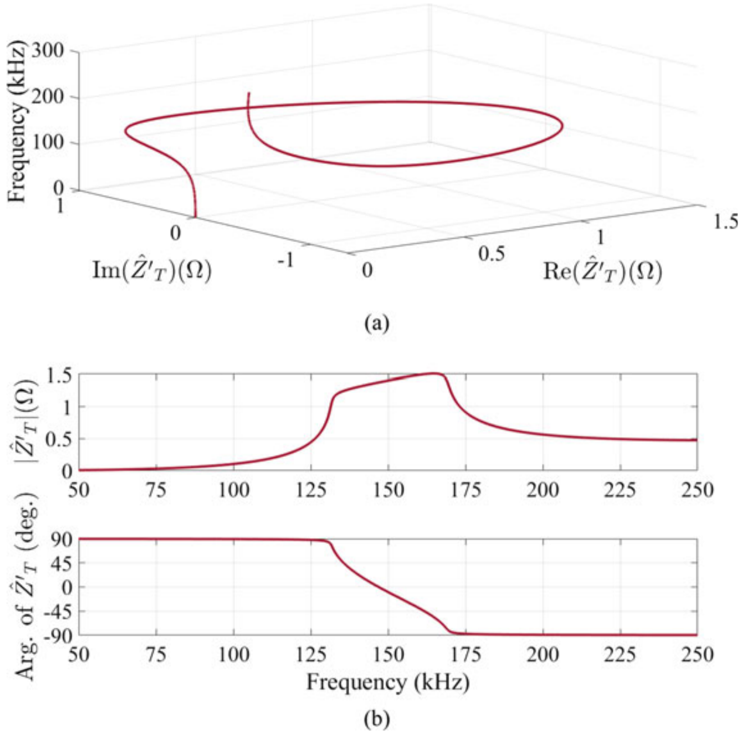


Fig. 5 Value of the impedance $\hat{Z}'_T = \frac{1}{2}\left(-\hat{Z} + \sqrt{\hat{Z}^2 + 4(\omega M)^2}\right)$ considering its (a) real and imaginary parts and (b) magnitude and argument for different values of the frequency

impedance $\hat{Z}_{eq}$ is constant regardless of the number of resonators. This value at the resonant frequency approximates to $\omega_0 M$ for $R \ll \omega_0 M$, as referred in [10].

Furthermore, when a receiver is placed over the $(l+1)$th resonator and the line is terminated with an impedance $\hat{Z}'_T$ equal to (27) we can determine $\hat{Z}_{eq}$ with (25) by setting $\hat{Z}'_T$ equal to $\frac{1}{2}\left(-\hat{Z} + \sqrt{\hat{Z}^2 + 4(\omega M)^2}\right) + \hat{Z}_d$ and replacing n with $n - l$.

Furthermore, using the values in [4], we can plot the variation of the impedance $\hat{Z}'_T = \frac{1}{2}\left(-\hat{Z} + \sqrt{\hat{Z}^2 + 4(\omega M)^2}\right)$ versus frequency as shown in Fig. 5.

4.3 Numerical Analysis of the Monotonicity of the Value of the Impedance

As seen in Sect. 3.4, the monotonicity of the odd or even terms of the sequence depends on the sign of $a_2 b_1 - a_1 b_2$, which, for the resonance frequency ($a = (\omega_0 M)^2$, $b = R$, $x_0 = R'_T$), is given by

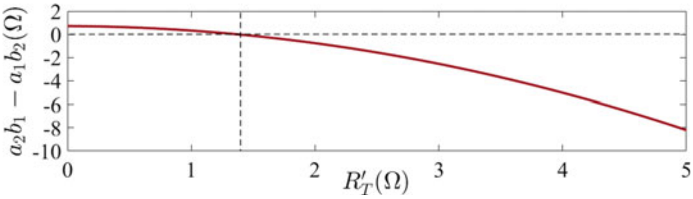


Fig. 6 Plot of $a_2b_1 - a_1b_2$ versus R'_T

$$\frac{(\omega_0 M)^2 - R'^2_T - RR'_T}{\sqrt{R^2 + 4(\omega_0 M)^2}} \tag{28}$$

Using the values determined experimentally in [4] the value of $a_2b_1 - a_1b_2$ is shown in Fig. 6.

It can be seen that $a_2b_1 - a_1b_2$ is zero for $R'_T = \frac{1}{2}\left(-R + \sqrt{R^2 + 4(\omega_0 M)^2}\right)$, positive for $R'_T < \frac{1}{2}\left(-R + \sqrt{R^2 + 4(\omega_0 M)^2}\right)$ and negative otherwise.

The values of the even and odd terms of the sequence of the continued fraction decrease and increase, respectively, for $R'_T < \frac{1}{2}\left(-R + \sqrt{R^2 + 4(\omega_0 M)^2}\right)$; otherwise, the opposite behaviour is observed.

4.4 Analysis of the Variation of the Speed of Convergence with the Variation of the Circuit Parameters

As in the previous section, it is assumed that the array is operating at the resonant frequency meaning that all the parameters become real. Recalling (21), defined in Sect. 3.5, for $n = 0$ (one resonator) and $n = 1$ (two resonators) we have

$$\delta_0 = \left| \frac{R + 2R'_T - \sqrt{R^2 + 4(\omega_0 M)^2}}{2} \right| \tag{29}$$

and

$$\delta_1 = \left| \frac{R - \sqrt{R^2 + 4(\omega_0 M)^2}}{2} + \frac{(\omega_0 M)^2}{R + R'_T} \right| \tag{30}$$

It can be seen in Fig. 7 that $\delta_0 = \delta_1 = 0$ for $R'_T = \frac{1}{2}\left(-R + \sqrt{R^2 + 4(\omega_0 M)^2}\right)$. Moreover, $\delta_0 < \delta_1$ when $R'_T < \frac{1}{2}\left(-R + \sqrt{R^2 + 4(\omega_0 M)^2}\right)$, and $\delta_0 > \delta_1$ for $R'_T > \frac{1}{2}\left(-R + \sqrt{R^2 + 4(\omega_0 M)^2}\right)$. The largest differences between the value of the equivalent impedance $\hat{Z}_{eq}$ and its limit (27) for $n \to \infty$, are given by δ_0 and δ_1.

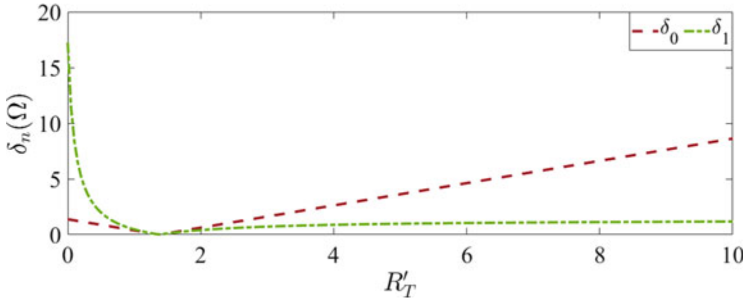


Fig. 7 Value of δ for different values of R'_T

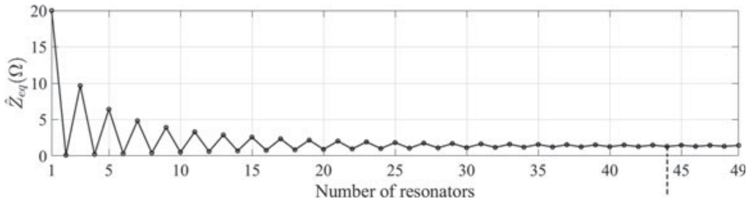


Fig. 8 Value of $\hat{Z}_{eq}$ versus the number of resonators, for $R'_T = 20\,\Omega$ and $\varepsilon = 0.7$

From (24), we have

$$N > \log_{\left|\frac{R-\sqrt{R^2+4(\omega_0 M)^2}}{R+\sqrt{R^2+4(\omega_0 M)^2}}\right|}\left(\frac{2}{R-\sqrt{R^2+4\left(\omega_0 M\right)^2}+2R'_T}\varepsilon\right), \tag{31}$$

for $R'_T > \frac{1}{2}\left(-R+\sqrt{R^2+4\left(\omega_0 M\right)^2}\right)$ or

$$N > \log_{\left|\frac{R-\sqrt{R^2+4(\omega_0 M)^2}}{R+\sqrt{R^2+4(\omega_0 M)^2}}\right|}\left(\frac{2\left(R+R'_T\right)}{2\left(\omega_0 M\right)^2-R'_T\left(\sqrt{R^2+4\left(\omega_0 M\right)^2}+R\right)}\varepsilon\right). \tag{32}$$

for $R'_T < \frac{1}{2}\left(-R+\sqrt{R^2+4\left(\omega_0 M\right)^2}\right)$.

The minimum number of resonators after the one connected to the voltage source so that the difference between $\hat{Z}_{eq}$ and $\lim_{n\to\infty}\hat{Z}_{eq}$ is within $\pm\varepsilon$ is given by the integer number N. N is calculated as the smallest integer greater than the logarithm of (31), when $R'_T > \frac{1}{2}\left(-R+\sqrt{R^2+4\left(\omega_0 M\right)^2}\right)$. As an example, for $R'_T = 20\,\Omega$ and $\varepsilon = 0.7$, which represents 51% of $\lim_{n\to\infty}\hat{Z}_{eq}$, the value of the logarithm is 42.7 and thus $N = 43$ (i.e. 44 resonators) as Fig. 8 shows.

5 Conclusion

In this paper, the theory of linear homogeneous difference equations is applied to the study of the continued fraction that represents the equivalent impedance of an array of resonators in order to obtain a more complete and rigorous understanding of the behaviour of the system. In this way, it is studied the explicit closed-form expression for the equivalent impedance which depends on the circuit parameters, termination impedance, number of resonators and position and impedance of receiver. Moreover, from the mathematical analysis of the convergence and monotonicity of the recursive sequence that defines the continued fraction, a better insight of the behaviour of the fraction with respect to its variables can be achieved. It is found that, for an arbitrarily large number of resonators, the equivalent impedance value is given only by the electrical parameters of the resonator array, $\hat{Z}$ and ωM, and that it does not depend on the value of the termination impedance, $\hat{Z}'_T$, and on the finite number of receivers over the line. Moreover, by terminating the resonator array with this impedance, the impedance of the resonator line is constant for any number of resonators. It is also proved that the recursive sequence used to model the system has an oscillating behaviour, having the even and odd subsequences opposite monotonicities (increasing or decreasing), according to the value of the termination impedance. As a consequence, for an array with a certain number of resonators, the equivalent impedance is bounded by the two first two terms of the recursive sequence (i.e., by the termination impedance, $\hat{Z}'_T$, and the equivalent impedance of one resonator seen from the cell connected to the source).

Appendix

Constants a_1,, a_2, b_1, c_1,, c_2, d_1, d_2 ,e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8:

$$a_1 = \frac{x_0}{2} - \frac{a}{\sqrt{b^2+4a}} + \frac{bx_0}{2\sqrt{b^2+4a}}; a_2 = \frac{x_0}{2} + \frac{a}{\sqrt{b^2+4a}} - \frac{bx_0}{2\sqrt{b^2+4a}}$$

$$b_1 = \frac{1}{2} - \frac{x_0}{\sqrt{b^2+4a}} - \frac{b}{2\sqrt{b^2+4a}}; b_2 = \frac{1}{2} + \frac{x_0}{\sqrt{b^2+4a}} + \frac{b}{2\sqrt{b^2+4a}}$$

$$c_1 = \frac{y_0}{2} - \frac{a}{\sqrt{b^2+4a}} + \frac{y_0 b}{2\sqrt{b^2+4a}}; c_2 = \frac{y_0}{2} + \frac{a}{\sqrt{b^2+4a}} - \frac{y_0 b}{2\sqrt{b^2+4a}}$$

$$d_1 = \frac{1}{2} - \frac{y_0}{\sqrt{b^2+4a}} - \frac{b}{2\sqrt{b^2+4a}}; d_2 = \frac{1}{2} + \frac{y_0}{\sqrt{b^2+4a}} + \frac{b}{2\sqrt{b^2+4a}}$$

$$e_1 = \hat{Z}_d\left(2\left(\omega M\right)^2 - f\hat{Z}_T\right); e_2 = \hat{Z}_d\left(2\left(\omega M\right)^2 - g\hat{Z}_T\right)$$

$$e_3 = g\left(\hat{Z} - \hat{Z}_d\right)\hat{Z}_T + 2\left(\omega M\right)^2\left(-h + \hat{Z}_d + 2\hat{Z}_T\right)$$

$$e_4 = f\left(\hat{Z} - \hat{Z}_d\right)\hat{Z}_T + 2\left(\omega M\right)^2\left(h + \hat{Z}_d + 2\hat{Z}_T\right)$$

$$e_5 = \left(\omega M\right)^2\left(4\left(\omega M\right)^2 + f\left(\hat{Z} + \hat{Z}_d\right) + 2\left(-h + \hat{Z}_d\right)\hat{Z}_T\right)$$

$$e_6 = \hat{Z}_d\left(f\hat{Z}\hat{Z}_T + \left(\omega M\right)^2\left(-f + 2\hat{Z}_T\right)\right); e_7 = \hat{Z}_d\left(g\hat{Z}\hat{Z}_T - \left(\omega M\right)^2\left(g - 2\hat{Z}_T\right)\right)$$

$$e_8 = \left(\omega M\right)^2\left(4\left(\omega M\right)^2 + g\left(\hat{Z} + \hat{Z}_d\right) + 2\left(h + \hat{Z}_d\right)\hat{Z}_T\right); h = \sqrt{\hat{Z}^2 + 4\left(\omega M\right)^2}$$

References

1. Ahmad, A., Alam, M.S., Chabaan, R.: A comprehensive review of wireless charging technologies for electric vehicles. IEEE Trans. Transport. Electrific. **4**(1), 38–63 (2018)
2. Alberto, J., Puccetti, G., Grandi, G., Reggiani, U., Sandrolini, L.: Experimental study on the termination impedance effects of a resonator array for inductive power transfer in the hundred kHz range. In: Proceedings of the IEEE Wireless Power Transfer Conference (WPTC 2015), Boulder, CO, USA, May 2015, pp. 1–4 (2015)
3. Alberto, J., Reggiani, U., Sandrolini, L.: Circuit model of a resonator array for a WPT system by means of a continued fraction. In: Proceedings of the IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow (RTSI), Bologna, Italy, September 2016, pp. 1–6 (2016)
4. Alberto, J., Reggiani, U., Sandrolini, L., Albuquerque, H.: Fast calculation and analysis of the equivalent impedance of a wireless power transfer system using an array of magnetically coupled resonators. Progress Electromagnet. Res. B **80**, 101–112 (2018)
5. Alberto, J., Reggiani, U., Sandrolini, L., Albuquerque, H.: Accurate calculation of the power transfer and efficiency in resonator arrays for inductive power transfer. Progress Electromagnet. Res. B **83**, 61–76 (2019)
6. Elaydi, S.: An Introduction to Difference Equations. Springer, New York (2005)
7. Liu, Z., Chen, Z., Guo, Y., Yu, Y.: A novel multi-coil magnetically-coupled resonance array for wireless power transfer system. In: 2016 IEEE Wireless Power Transfer Conference (WPTC), May 2016, pp. 1–3
8. Monti, G., Corchia, L., Tarricone, L., Mongiardo, M.: A network approach for wireless resonant energy links using relay resonators. IEEE Trans. Microw. Theory Tech. **64**(10), 3271–3279 (2016)
9. Ranum, B.T., Rahayu, N.W.D.E., Munir, A.: Development of wireless power transfer receiver for mobile device charging. In: The 2nd IEEE Conference on Power Engineering and Renewable Energy (ICPERE), December 2014, pp. 48–51 (2014)
10. Stevens, C.J.: Magnetoinductive waves and wireless power transfer. IEEE Trans. Power Electron. **30**(11), 6182–6190 (2015)
11. Xue, R.F., Cheng, K.W., Je, M.: High-efficiency wireless power transfer for biomedical implants by optimal resonant load transformation. IEEE Trans. Circ. Syst. I Reg. Pap. **60**(4), 867–874 (2013)
12. Zhang, X., Ho, S.L., Fu, W.N.: Quantitative design and analysis of relay resonators in wireless power transfer system. IEEE Trans. Magn. **48**(11), 4026–4029 (2012)
13. Zhong, W.X., Lee, C.K., Hui, S.Y.: Wireless power domino-resonator systems with noncoaxial axes and circular structures. IEEE Trans. Power Electron. **27**(11), 4750–4762 (2012)
14. Alberto, J.M.M.A.V.: Mathematical approach for an accurate solution of the circuit model of resonator arrays for inductive power transfer. Ph.D. dissertation, Alma, Aprile 2017

Asymptotics of an Equation with Large State-Dependent Delay

Ilia Kashchenko

Abstract In this paper, the local dynamics and asymptotic approximation of solutions of equation with large state-dependent delay is studied. Stability of equilibrium is studied and critical cases are identified. When values of parameters are close to critical ones the nonlinear parabolic equations are constructed. Solutions of these equations determine the behaviour of the solutions and main terms of asymptotics of the solution.

1 Introduction

Delay differential equations are used as mathematical models of various physical and biological systems [4, 7, 17, 18, 20]. The phase space of DDEs is infinite-dimensional [8]. This specific feature of delay systems not only makes them more difficult to analysis, but also complicates the dynamics [6, 12, 25].

This article is devoted to the asymptotic of solutions of state-dependent delay differential equations (SD-DDE). In these equations delay is not constant, but depends on the state of the system. Problems of such kind arise in various applications (see, e.g., [3, 11, 22, 23, 26]). Some results about existence and stability of solutions (especially periodic solutions) of SD-DDE one may found, for example, in [2, 5, 9, 10, 19, 21].

This paper extends previous article [15]. Here we will study the situation when time delay is a given function of the state of the system not only in current moment, but also in previous moments. At the same time important assumption is that the delay time is sufficiently large. The same assumption in the systems with constant delay leads to a complicated and high-dimensional dynamics [1, 13, 24, 25].

Consider the nonlinear differential-difference equation with state-dependent delay

$$\dot{u} + u = F(u(t - T\varphi(u_h))), \quad u_h = \int\limits_{-h}^{0} u(t+s)dr(s) \quad (\int\limits_{-h}^{0} dr(s) = 1) \qquad (1)$$

I. Kashchenko (✉)
P. G. Demidov Yaroslavl State University, Sovetskaya str. 14, Yaroslavl, Russia
e-mail: iliyask@uniyar.ac.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_26

where $F(u)$ and $\varphi(u) > 0$ are sufficiently smooth functions and $T, h > 0$ are a parameters. The main assumption is that the delay parameter T is sufficiently large: $T \gg 1$.

Assume $F(0) = 0$, so Eq. (1) has equilibrium $u \equiv 0$. Let study the behaviour of solutions of Eq. (1) from small fixed (i.e., T-independent) neighbourhood of the zero equilibrium.

Introduce the phase space. Let r_1, r_2 and R are positive numbers such that $|\varphi(v)| \leq R$ for all $|v| \leq r_2$ and $|u_h| \leq r_2$ for all u such that $\max |u| \leq r_1$. For a sufficiently small fixed r_0 let the initial data for solutions in (1) belong to $C_{[-TR(r_0),\,0]}$.

Represent non-linear functions $F(v)$ and $\varphi(v)$ in a neighbourhood of zero in the form

$$F(v) = av + bv^2 + cv^3 + O(v^4),$$
$$\varphi(v) = 1 + \alpha v + \beta v^2 + O(v^3).$$

Using normalization $t \to Tt$ in (1) we obtain the equation

$$\varepsilon\dot{u} + u = F(u(t - \varphi(u_{\varepsilon,h})), \quad u_{\varepsilon,h} = \int\limits_{-h}^{0} u(t + \varepsilon s)dr(s), \quad 0 < \varepsilon = T^{-1} \ll 1. \tag{2}$$

Now we have the problem to study the behaviour of solutions of Eq. (2) from small (but ε-independent) neighbourhood of zero in the space $C_{[-R(r_0),\,0]}$.

In the case $\varphi(v) \equiv 1$ this problem was studied in [14, 16]. It was shown that the critical cases have an infinite dimension (i.e., it is unbounded as $\varepsilon \to 0$). As the main results, special nonlinear boundary value problems of the parabolic type were constructed which nonlocal dynamics determine the behaviour of the solutions to Eq. (2) (for small ε) that belong to a sufficiently small ε-independent neighborhood of zero. Below, the technique from [14, 16], based on the construction of quasi-normal forms is used to study Eq. (2). Specifically, it will be shown that the dynamics of this equation can be more complicated than in the case of a constant delay.

2 Linear Analysis

Let us start with a linear analysis. Linearised at the zero equilibrium equation for Eq. (2) is equation with constant delay

$$\varepsilon\dot{x} + x = ax(t - 1).$$

It's the same as in the case $\varphi(v) \equiv 1$ (see [14, 16]). Thus, characteristic equation

$$\varepsilon\lambda + 1 = a\exp(-\lambda) \tag{3}$$

is the same as in the case of constant delay. Location of its roots determines the stability of equilibrium [8].

From [14, 16] we have the next result.

Theorem 1. *For $|a| < 1$ all roots of (3) have negative real parts that are separated from the imaginary axis as $\varepsilon \to 0$. All solutions of Eq. (2) from a sufficiently small ε-independent neighborhood of $u = 0$ tend to zero as $t \to \infty$.*

If $|a| > 1$, then there is a root of (3), such that its real part is positive and separated from zero as $\varepsilon \to 0$. Zero equilibrium is unstable and the dynamics of (2) became non-local: there are no stable regimes in neighbourhood of $u = 0$.

From this Theorem it follows that it is necessary to consider the case when the parameter a is close to ± 1.

Further, consider the nearly critical case

$$a = \pm(1 + \varepsilon^p a_1), \qquad p > 0.$$

Under these conditions, the characteristic Eq. (3) has infinitely many roots with real parts tending to zero as $\varepsilon \to 0$. By applying methods based on the construction of normal forms, the analysis of the local dynamics of (2) in the first approximation will be reduced to study of special nonlinear parabolic equation—quasinormal form [14, 16].

Consider cases $a \approx 1$ and $a \approx -1$ separately.

3 Critical Case $a \approx 1$

First let $a = 1 + \varepsilon^2 a_1$. In this case an infinite number of roots of (3) tend to $2k\pi i$ ($k \in Z$). Consider asymptotic series

$$u = \varepsilon^2 \xi(\tau, x) + \varepsilon^3 u_2(\tau, x) + \cdots , \tag{4}$$

where $\tau = \varepsilon^2 t$, $x = (1 - \varepsilon + \varepsilon^2)t$, and the dependence on x is 1-periodic. Substituting (4) into (2) and collecting same powers of ε, at the second step the solvability condition for the resulting equation for u_2, yields an equation for determining the unknown $\xi(\tau, x)$:

$$\frac{\partial \xi}{\partial \tau} = \frac{1}{2}\frac{\partial^2 \xi}{\partial x^2} + a_1 \xi + b\xi^2 - \alpha \xi \frac{\partial \xi}{\partial x} \tag{5}$$

with periodic boundary conditions

$$\xi(\tau, x + 1) \equiv \xi(\tau, x). \tag{6}$$

Note, that in contrast to the case of constant delay ($\alpha = 0$) the nonlinearity in this system contains the derivative of ξ with respect to x.

Solutions of (5), (6) give a zero-order approximation for solutions of (2).

Theorem 2. *Let the boundary value problem (5), (6) has a solution $\xi_0(\tau, x)$ bounded as well as its derivatives. Then*

$$u_0(t, \varepsilon) = \varepsilon^2\xi_0(\varepsilon^2 t, (1-\varepsilon+\varepsilon^2)t) \tag{7}$$

produce in Eq. (2) asymptotically small discrepancy of order $o(\varepsilon^4)$ uniformly for all $t \in [0, +\infty)$.

Proof. Denote $\tau = \varepsilon^2 t, x = (1-\varepsilon+\varepsilon^2)t$. Using periodicity of $xi_0(\tau, x)$ and Taylor series expansions we obtain asymptotic formulas

$$\begin{aligned}
u_{\varepsilon,h} &= \varepsilon^2\xi_0(\tau, x) + o(\varepsilon^2),\\
\varphi(u_{\varepsilon,h}) &= 1 + \varepsilon^2\alpha\xi_0(\tau, x) + o(\varepsilon^2),\\
u(t-\varphi(u_{\varepsilon,h})) &= \varepsilon^2\xi_0 + \varepsilon^3\frac{\partial\xi_0}{\partial x} - \varepsilon^4\frac{\partial\xi_0}{\partial\tau} - \varepsilon^4\frac{\partial\xi_0}{\partial x} - \varepsilon^4\alpha\xi_0\frac{\partial\xi_0}{\partial x} + \frac{1}{2}\varepsilon^4\frac{\partial^2\xi_0}{\partial x^2} + o(\varepsilon^4),\\
\dot{u} &= \varepsilon^2(1-\varepsilon)\frac{\partial\xi_0}{\partial x} + o(\varepsilon^3).
\end{aligned}$$

Here all residuals $o(\varepsilon^2)$, $o(\varepsilon^3)$ and $o(\varepsilon^4)$ are uniform due to the fact that $\xi_0(\tau, x)$ and its derivatives are bounded.

Using these formulas substitute (7) into Eq. (2) and obtain that all terms which are bigger than $O(\varepsilon^4)$ are vanishes and we have

$$o(\varepsilon^4) = -\varepsilon^4\frac{\partial\xi_0}{\partial\tau} - \varepsilon^4\alpha\xi_0\frac{\partial\xi_0}{\partial x} + \frac{1}{2}\varepsilon^4\frac{\partial^2\xi_0}{\partial x^2} + \varepsilon^4 a_1\xi_0 + \varepsilon^4 b\xi_o^2 + o(\varepsilon^4).$$

Since $\xi_0(\tau, r)$ is a solution of (5), the discrepancy is small of order of $o(\varepsilon^4)$ uniformly for all $t \in [0, +\infty)$.

When $a = 1 + \varepsilon^p a_1$ $(0 < p < 2)$, the situation is much more complicated. In contrast to the previous case, steady states are formed at asymptotically high (as $\varepsilon \to 0$) "modes", i.e. the second argument of ξ is rapidly oscillating. Also, the order of the solution amplitude varies as well.

Introduce the following notation. Let $z > 0$ is arbitrary fixed and let $\Theta_z = \Theta_z(\varepsilon) \in [0, 1)$ produce an integer when added to $\varepsilon^{p/2-1}z$. Function $\Theta_z(\varepsilon)$ is bounded, piecewise continues and take all its values infinite number of times when $\varepsilon \to 0$.

An analogue of (4) is the series

$$u = \varepsilon^p\xi(\tau, x) + \varepsilon^{2p}u_2(\tau, x) + \cdots,$$

where $\tau = \varepsilon^p t$ and $x = (\varepsilon^{p/2-1}z + \Theta_z - \varepsilon^{p/2}z)t$. Note that $p/2 - 1 < 0$, so x is a "fast" time.

Then we obtain the following equation for determining $\xi(t, x)$

$$\frac{\partial \xi}{\partial \tau} = \frac{1}{2} z^2 \frac{\partial^2 \xi}{\partial x^2} + a_1 \xi - \alpha z \xi \frac{\partial \xi}{\partial x} \tag{8}$$

with periodic boundary conditions (6).

Note that (8) contain an arbitrary parameter z at the second derivative with respect to r and in nonlinearity.

Theorem 3. *Let $z > 0$, $\xi_0(\tau, x)$ be solution of (8), (6) bounded as well as its derivatives for all $\tau \geq 0$ and $x \in [0, 1]$. Then*

$$u_0(t, \varepsilon) = \varepsilon^p \xi(\varepsilon^p t, (\varepsilon^{p/2-1} z + \Theta_z - z\varepsilon^{p/2})t)$$

produce in (2) asymptotically small discrepancy of order $O(\varepsilon^{2p})$ uniformly for all $t \geq 0$.

Proof. Proof of this theorem is similar to the proof of Theorem 2.

4 Critical Case $a \approx -1$

Now let $a = -1 + \varepsilon^2 a_1$. Then (3) has infinitely many roots tending to $\pi(2k + 1)i$ ($k = 0, \pm 1, \pm 2, \ldots$). Consider the asymptotic series

$$u = \varepsilon \xi(\tau, x) + \varepsilon^2 u_2(\tau, x) + \varepsilon^3 u_3(\tau, x) + \cdots, \tag{9}$$

where $\tau = \varepsilon^2 t$, $x = (1 - \varepsilon + \varepsilon^2)t$, functions $u_2(\tau, x)$ and $u_3(\tau, x)$ are periodic by x with the period equal to 1, and $\xi(\tau, x)$ is the linear combination of the critical modes $\exp(\pi(2k + 1)ix)$:

$$\xi(\tau, x) = \sum_{k=-\infty}^{\infty} \xi_k(\tau) \exp(\pi(2k + 1)ix).$$

Substitute (9) into (2) and collect same powers of ε. At ε^1 we get identity, then at ε^2 we obtain

$$u_2 = \frac{1}{2}(b\xi^2 - \alpha\xi \frac{\partial \xi}{\partial x}).$$

Finally, after simplifying the equality at ε^3 we get complex parabolic equation for determining $\xi(\tau, x)$:

$$\frac{\partial \xi}{\partial \tau} = \frac{1}{2} \frac{\partial^2 \xi}{\partial x^2} - a_1 \xi - (b^2 + c)\xi^3 + (\frac{1}{2}\alpha b - \beta)\xi^2 \frac{\partial \xi}{\partial x} - \frac{1}{2}\alpha^2 \xi (\frac{\partial \xi}{\partial x})^2 \tag{10}$$

with anti-periodic boundary conditions

$$\xi(\tau, x+1) = -\xi(\tau, x). \tag{11}$$

Theorem 4. *Let the boundary value problem (10), (11) has a solution $\xi_0(\tau, x)$ bounded as well as its derivatives for all $\tau \geq 0$ and $x \in [0, 1]$. Then*

$$u_0(t, \varepsilon) = \varepsilon\xi_0(\tau, x) + \frac{1}{2}\varepsilon^2(b\xi_0^2(\tau, x) - \alpha\xi_0\big(\tau, x\big)\frac{\partial\xi_0(\tau, x)}{\partial x})$$

produce in Eq. (2) an asymptotically small discrepancy of order $O(\varepsilon^3)$ uniformly for all $t \geq 0$. where $\tau = \varepsilon^2 t$ and $x = (1 - \varepsilon + \varepsilon^2)t$.

Proof. This theorem can be proved by the same way as Theorem 2.

Note that, in some cases, solutions of problem (10), (11) can be found analytically. For example, consider a standard bifurcation from the equilibrium of Eq. (10) with $a_1 = -\frac{1}{2}\pi^2 + \mu, \quad |\mu| \ll 1$. Here, we have a double zero root at the origin (in the linear problem) with two groups of solutions.

Under the condition $a_1 = -1 + \varepsilon^p a_1$, where $0 < p < 2$, the situation is more complicated.

Let us formulate the result. Values z and Θ_z have the same meaning as before.

Consider the boundary value problem

$$\frac{\partial\xi}{\partial\tau} = \frac{z^2}{2}(1 - \alpha^2\xi^2)\frac{\partial^2\xi}{\partial x^2} - a_1\xi - (b^2 + c)\xi^3 + (\frac{1}{2}\alpha b - \beta)z\xi^2\frac{\partial\xi}{\partial x} - \frac{z^2}{2}\alpha^2\xi(\frac{\partial\xi}{\partial x})^2. \tag{12}$$

with a boundary condition

$$\xi(\tau, x) = -\xi(\tau, x+1). \tag{13}$$

Theorem 5. *Let for some $z > 0$ Eq. (12) has solution $\xi_0(\tau, x)$ bounded for all $\tau \geq 0$ as well as its derivatives. Then*

$$u_0(t, \varepsilon) = \varepsilon^{p/2}\xi_0(\tau, x) + \frac{1}{2}\varepsilon^p(b\xi_0^2(\tau, x) - \alpha\xi_0\big(\tau, x\big)\frac{\partial\xi_0(\tau, x)}{\partial x})$$

produce in (2) asymptotically small discrepancy of order $o(\varepsilon^p)$ uniformly for all $t \geq 0$. Here $\tau = \varepsilon^p t$ and $x = (\varepsilon^{\frac{p}{2}-1}z + \Theta_z - z\varepsilon^{\frac{p}{2}})t$.

5 Conclusion

The differential equation with large state-dependent delay was considered. We have identified critical cases in the equilibrium stability problem. In critical cases special nonlinear boundary value problems was constructed. Its solutions determine the main terms of asymptotic approximations of solutions to the original SD-DDE in a small neighbourhood of the equilibrium. Very important that these boundary value problems does not depend on small parameter and easily may be solved numerically.

The coefficients of constructed boundary value problems depend on the parameters α and β of the function $\varphi(v)$ and does not depend on the function $r(s)$ from definition of u_h. Thus main terms of solutions and their dynamics does not depend on $r(s)$ too.

Acknowledgements This work was funded by RFBR according to the research project 18-29-10043.

References

1. Bestehorn, M., Grigorieva, E.V., Haken, H., Kashchenko, S.A.: Order parameters for class-b lasers with a long time delayed feedback. Phys. D **145**(1–2), 110–129 (2000). https://doi.org/10.1016/S0167-2789(00)00106-8
2. Brokate, M., Colonius, F.: Linearizing equations with state-dependent delays. Appl. Math. Optim. **21**(1), 45–52 (1990)
3. Crocco, L., Harrje, D.T., Reardon, F.H.: Transverse combustion instability in liquid propellant rocket motors. ARS J. **32**(3), 366–373 (1962)
4. Erneux, T.: Applied Delay Differential Equations. Springer, Berlin (2009)
5. Golubenets, V.: Local bifurcations analysis of a state-dependent delay differential equation. Autom. Control Comput. Sci. **50**(7), 617–624 (2016)
6. Grigorieva, E., Kaschenko, S.: Stability of equilibrium state in a laser with rapidly oscillating delay feedback. Phys. D **291**, 1–7 (2015)
7. Haken, H.: Brain Dynamics: Synchronization and Activity Patterns in Pulse-Coupled Neural Nets with Delays and Noise. Springer, Berlin (2002)
8. Hale, J., Sjoerd, M.: Introdution to Functional Differential Equations. Springer, New York (1993)
9. Hartung, F., Turi, J.: On differentiability of solutions with respect to parameters in state-dependent delay equations. J. Differ. Equ. **135**(2), 192–237 (1997)
10. Hu, Q., Wu, J.: Global hopf bifurcation for differential equations with state-dependent delay. J. Differ. Equ. **248**(12), 2801–2840 (2010)
11. Insperger, T., Barton, D.A., Stépán, G.: Criticality of hopf bifurcation in state-dependent delay model of turning processes. Int. J. Non-Linear Mech. **43**(2), 140–149 (2008)
12. Kashchenko, A.: Multistability in a system of two coupled oscillators with delayed feedback. J. Differ. Equ. **266**(1), 562–579 (2019)
13. Kashchenko, I., Kaschenko, S.: Normal and quasinormal forms for systems of difference and differential-difference equations. Commun. Nonlinear Sci. Numer. Simul. **38**, 243–256 (2016). https://doi.org/10.1016/j.cnsns.2016.02.041
14. Kashchenko, I.S.: Local dynamics of equations with large delay. Comput. Math. Math. Phys. **48**(12), 2172–2181 (2008)
15. Kashchenko, I.S., Kashchenko, S.A.: Local dynamics of an equation with a large state-dependent delay. Dokl. Math. **92**(2), 581–584 (2015). https://doi.org/10.1134/S1064562415050221
16. Kashchenko, S.A.: Application of the normalization method to the study of the dynamics of a differential-difference equation with a small factor multiplying the derivative. Differ. Uravneniya **25**(8), 1448–1451 (1989)

17. Kashchenko, S.A.: Models of Wave Memory. Lecture Notes in Morphogenesis. Springer (2015). https://doi.org/10.1007/978-3-319-19866-8
18. Kolmanovskii, V., Myshkis, A.: Introduction to the Theory and Applications of Functional Differential Equations. Springer, Berlin (2013)
19. Krisztin, T., Walther, H.O.: Smoothness issues in differential equations with state-dependent delay. Rend. Istit. Mat. Univ. Trieste: Int. J. Math. **49**, 95–112 (2017)
20. Mackey, M.C., Glass, L.: Oscillation and chaos in physiological control systems. Science **197**(4300), 287–289 (1977)
21. Mallet-Paret, J., Nussbaum, R.D., Paraskevopoulos, P.: Periodic solutions for functional differential equations with multiple state-depend time lags. Topol. Methods Nonlinear Anal. **3**(1), 101–162 (1994)
22. Reardon, F.H., Crocco, L., Harrje, D.T.: Velocity effects in transverse mode liquid propellant rocket combustion instability. AIAA J. **2**(9), 1631–1641 (1964)
23. Sabersky, R.H.: Effect of wave propagation in feed lines on low frequency rocket instability. Jet Propul. **24**, 172–174 (1954)
24. Wolfrum, M., Yanchuk, S.: Eckhaus instability in systems with large delay. Phys. Rev. Lett. **96**, 220,201 (2006)
25. Yanchuk, S., Perlikowski, P.: Delay and periodicity. Phys. Rev. E **79**, 046,221 (2009)
26. Zager, M.G., Schlosser, P.M., Tran, H.T.: A delayed nonlinear PBPK model for genistein dosimetry in rats. Bull. Math. Biol. **69**(1), 93 (2007)

About Some Methods of Analytic Representation and Classification of a Wide Set of Geometric Figures with “Complex” Configuration

Ilia Tavkhelidze, J. Gielis, and S. Pinelas

Abstract We will present 2 different analytical representations of only one general idea—this is the representation of complex movements using the superposition of certain elementary displacements! Despite of the analytical and structural similarity of these representations, they describe fundamentally different geometric figures (in statics) and trajectories of motion (in dynamics). In previous articles [1–9] a wide class of geometric figures—“Generalized Twisting and Rotated” bodies GRT_m^n in short—was defined through their analytic representation. In particular cases, this analytic representation gives back many classical objects (torus, helicoid, helix, Möbius strip ... etc.). The aim of this article is to consider some geometric properties of a wide subclass of the generally defined surfaces. We show some geometric properties of GRT and GML—surfaces.

Keywords Analytic representation · Möbius strip · Möbius-Listing’s surfaces · Jacobi matrix · Jacobian · One-to one transformation · Regular points · Self-cross points

2000 MSC 53A05 · 51B10 · 57M25

I. Notations and Abbreviations. In this article we use the following notations:

- X, Y, Z and x, y, z, t—is the ordinary notation for space and time coordinates;
- τ, ψ, θ—are space values (local coordinates or parameters in parallelogram)

I. Tavkhelidze (✉)
Faculty of Exact and Natural Sciences, Tbilisi State University,
University Street 13, 0186 Tbilisi, Georgia
e-mail: ilia.tavkhelidze@tsu.ge

J. Gielis (✉)
Department of Biosciences Engineering, University of Antwerp, Antwerp, Belgium
e-mail: johan.gielis@uantwerpen.be

S. Pinelas (✉)
Department of Exact and Natural Sciences, Military Academy,
Av. Conde Castro Guimarães, 720-113 Amadora, Portugal
e-mail: sandra.pinelas@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_27

where:

$$\begin{aligned}&1.\ \tau \in [\tau_*, \tau^*], \text{ where } \tau_* \le \tau^* \ \mathit{usually\ are\ non\text{-}negative\ constants};\\ &2.\ \psi \in [0, 2\pi];\\ &3.\ \theta \in [0, 2\pi h], \quad \text{where} \ \ h \in \mathbf{R}\ (Real);\end{aligned} \tag{1}$$

But sometimes, as a special case, we suppose that

$$\tau \in [-\tau^*, \tau^*] \tag{1*}$$

- $P_m \equiv A_1A_2 \ldots A_m$ denotes a "Plane figure with m-symmetry", in particular P_m is a "regular polygon" or a star polygon, and m is the number of its angles or vertices. In the general case the edges of "regular polygons" are not always straight lines (A_iA_{i+1} may be, for example: edge of epicycloid, or edge of hypocycloid, or part of lemniscate of Bernoulli, and so on).
- $PR_m \equiv A_1A_2 \ldots A_mA'_1A'_2 \ldots A'_m$ denotes an orthogonal prism, whose ends $A_1A_2 \ldots A_m$ and $A'_1A'_2 \ldots A'_m$ are "Plane m-symmetric figures" P_m;

For example:

– PR_0—is a segment and P_0 is a point;

– PR_1—is an orthogonal cylinder, whose cross section is a P_1—plane figure without symmetry;

– $PR_2 \equiv A_1A_2A'_1A'_2$ is a rectangle, if $P_2 \equiv A_1A_2$ is a segment of straight line; but also PR_2 may be a cylinder with cross section P_2 (ellipse, or lemniscate of Bernoulli and so on);

– PR_∞—is an orthogonal cylinder, whose cross section is a P_∞-circle.

$$x = p(\tau, \psi, t); \quad z = q(\tau, \psi, t); \tag{2}$$

or

$$x = p(\tau, \psi, t)\cos\psi; \quad z = p(\tau, \psi, t)\sin\psi; \tag{2*}$$

are the analytic representations of the shape of a **"Plane figure with m-symmetry"** P_m which may change over time, but $p(0, 0, t) = q(0, 0, t) = 0$ and the point $(0, 0, t)$ is always the center of symmetry of this polygon (see [2, 7]).

- $D(p, q, t)$ or $D(p, t)$—diameter of plane figure P_m;
- OO'—axis of symmetry of the prism PR_m;
- $g(\theta)$—an arbitrary sufficiently smooth function

$$g(\theta) : [0, 2\pi h] \to [0, 2\pi h]; \tag{3}$$

and if $h = 1$, then for every $\Theta \in [0, 2\pi]$ there exists $\theta \in [0, 2\pi]$, such that $\Theta = g(\theta)$;

- $mod_m(n)$-natural number $< m$; for every two numbers $m \in \mathbf{N}$ (natural) and $n \in \mathbf{Z}$ (integer) there exists a unique representation $n = km + j \equiv km + mod_m(n)$, where $k \in \mathbf{Z}$ and $j \equiv mod_m(n) \in \mathbf{N} \bigcup \{0\}$;

$$\mu \equiv \begin{cases} n/m, & when\ m \in \mathbf{N} \quad n \in \mathbf{Z} \\ n, & when\ m = \infty\ n \in \mathbf{Z}\ or\ (n \in \mathbf{R}\ (Real)) \end{cases} \tag{4}$$

IIA. Generalized Twisting and Rotated bodies in dynamics—GTR_m^n (sometimes called **"Surfaces of revolution"** see [7, 11]) are defined by the parametric representations:

$$\begin{aligned} X(\tau, \psi, \theta, t) &= T_1(t) + [R(\psi, \theta, t) + p(\tau, \psi, \theta, t)\cos(\psi + \mu g(\theta))]\cos(\theta + M(t)) \\ Y(\tau, \psi, \theta, t) &= T_2(t) + [R(\psi, \theta, t) + p(\tau, \psi, \theta, t)\cos(\psi + \mu g(\theta))]\sin(\theta + M(t)) \\ Z(\tau, \psi, \theta, t) &= T_3(t) + Q(\theta, t) + p(\tau, \psi, \theta, t)\sin(\psi + \mu g(\theta)) \end{aligned} \tag{5}$$

where, respectively:

– the arguments (τ, ψ, θ, t) are defined in (1);

– the functions $R(\psi, \theta, t)\cos(\theta + M(t))$ and $R(\psi, \theta, t)\sin(\theta + M(t))$ define the **"Shape of the plane basic line"**, more precisely "Shape of the orthogonal projection on the plane XOY of the basic line" of corresponding body (see e.g.: circle in Figs. 1a, b, c, g; ellipse in Fig. 1e; spiral in Figs. 1d, f, i and square in Fig. 1h);

1. If $R(\psi, \theta, t) = const. > 0$ the basic line is a circle with radius R;

2. In general this may be any plane curve for example $R(\psi, \theta, t) = \varrho(\theta)$ Gielis curve (6) [10]

$$\varrho(\theta) = \left(\left| \frac{\cos(\frac{m_1}{4} \cdot \theta)}{A} \right|^{n_2} + \left| \frac{\sin(\frac{m_2}{4} \cdot \theta)}{B} \right|^{n_3} \right)^{-\frac{1}{n_1}} \tag{6}$$

– Function $p(\tau, \psi, \theta, t)$ defines the **"Shape of the radial cross section"** of the corresponding figure.

1. If $p(\tau, \psi, \theta, t) = \tau$ defined by (1*) and (5) (where $m = 2;\ n = 1;\ g(\theta) \equiv \theta$) this figure is a classic Möbius strip;

2. In general case this radius $p(\tau, \psi, \theta, t)$ may vary depending on angle ψ (this means for fixed values θ and t, the radial cross sections have different shapes, for example Gielis curve (6) $p(\tau, \psi, \theta, t) = \varrho(\psi)$ with different parameters, e.g. a half-angle for $m_1 = m_2 = 1/2$ and argument ψ instead θ in (6)), and on angle θ, but also over time (some examples of figures with different radial cross sections are shown: epicycloid $k = 6$ in Fig. 1: b and h; hypocycloid $k = 4$ in Fig. 1c; trifolium curve in Fig. 1g; square in Fig. 1e and f; circle in Fig. 1d; variable radius ellipse in Fig. 1i).

The shape of the plane basic line and of the shape of the radial cross section are the two fundamental shapes, perpendicular to each other. A special case is the torus, with the smaller and the larger circle defining the torus. The ratio of smaller to larger circle allows for a continuous transformation from sphere to torus, with spindle tori as intermediate shapes.

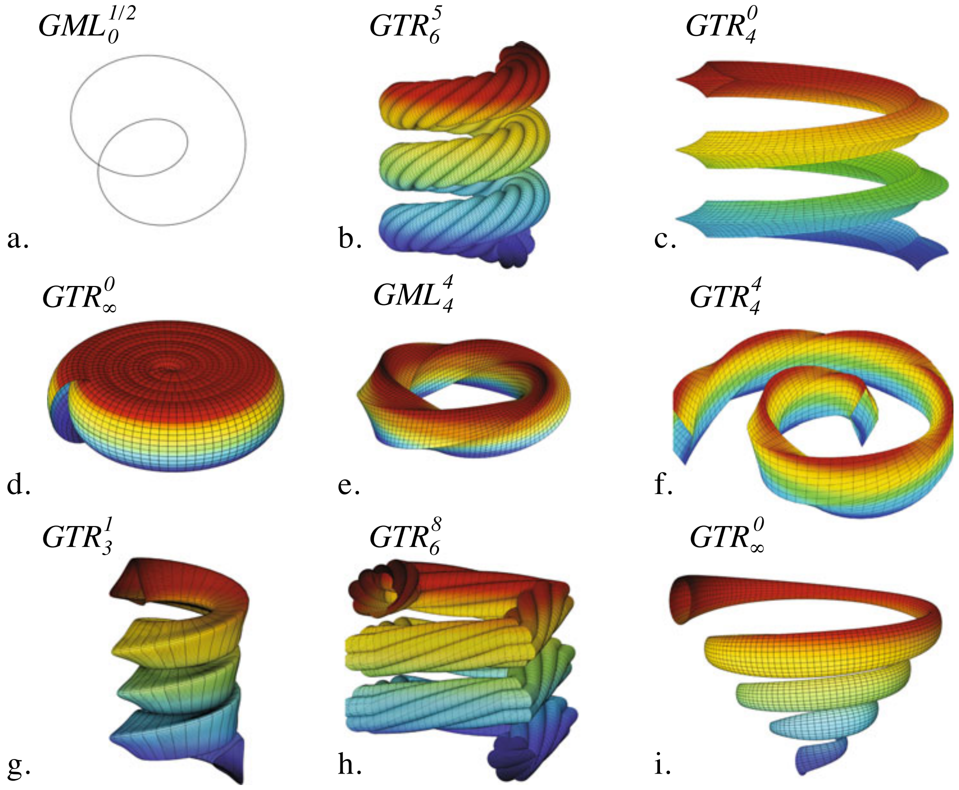


Fig. 1 Some examples of GRT_m^n lines, surfaces and bodies

When the shape of the radial cross section describes the curve only, a surface is described, but when also the disk of the radial cross section is included (e.g. when in (5) and (6) max $p(\tau, \psi, \theta, t) = \max \varrho(\psi) \equiv D(p, t) \leq \min R(\psi, \theta)$ is used), the results are bodies. In the same way shells can be defined for $\tau \in [\tau_*, \tau^*]$ (where $\tau^* - \tau_*$ is sufficiently small) a restricted range.

Additional functions are:

– The function $g(\theta)$ from (3) defines the **"Rule of twisting around basic line"**. For any function $g(\theta)$ the movement is called **semi-regular**; it is called **regular** for $g(\theta) \equiv \theta$.

– The number μ in (4) defines the **"Characteristic of twisting"**;

– $Q(\theta, t)$ is a smooth function which defines the **"Law of vertical stretching of figure"** (see e.g.: $Q \equiv 0$ in Fig. 1d,e,f; $Q \equiv const. \neq 0$ in Fig. 1b,c,g; $\partial Q/\partial\theta \neq 0$ in Fig. 1h,i);

– $(T_1(t); T_2(t); T_3(t))$—vector of displacement of a given body in space as a whole.

Therefore, this parametric representation defines a GTR_m^n body (some examples are shown in Fig. 1) with the following restrictions:

(1) The OO'-axis of symmetry (middle line) of the prism PR_m is transformed into a **"Basic line"** (sometimes called **"Profile curve"**);

(2) Rotation at the end of the prism (2) or (2*) is semi-regular along the middle line OO', or the twisting of the shape of radial cross section around the basic line is semi-regular (depending on $g(\theta)$).

IIB Generalized Möbius—Listing's Bodies— in short $GML_m^n\{\mu\}$ ($\mu \in \mathbf{Q}$ defined in (4))—is obtained by identifying the opposite ends of the prism PR_m in such a way that:

(A) For any integer $n \in \mathbf{Z}$ and $i = 1, \cdots, m$ each vertex A_i coincides with $A'_{(i+n)} \equiv A'_{(mod_m(i+n))}$, and each edge $A_i A_{(i+1)}$ coincides with the edge

$$A'_{(i+n)} A'_{(i+n+1)} \equiv A'_{(mod_m(i+n))} A'_{(mod_m(i+n+1))}$$

correspondingly;

(B) The integer $n \in \mathbf{Z}$ denotes the number of rotations of the end of the prism with respect to the axis OO' before the identification. If $n > 0$, the rotations are counterclockwise, and if $n < 0$ then rotations are clockwise. Some particular examples of GML_m^n and its graphical realizations can be found in [2, 4, 6–8] (see e.g. Fig. 1e, 2 or 3).

Definition 1. The Basic line of the $GML_m^n\{\mu\}$ body for each $\mu \in \mathbf{Q}$, is a continuous closed plane or spatial line on which the axis of symmetry OO' of the prism PR_m transforms after identifying the ends of the prism. (examples of basic lines: Fig. 2 circle, i.e. $\mu = 0$, Fig. 3d torus line $\mu = 3$ [6]).

Definition 2. A Rib of the $GML_m^n\{\mu\}$ body for each $\mu \in \mathbf{Q}$, is a continuous closed line, in which only the vertices of the radial cross sections (plane m-symmetric figures) of this body or ribs of prism PR_m are situated; i.e. torus line with characteristic

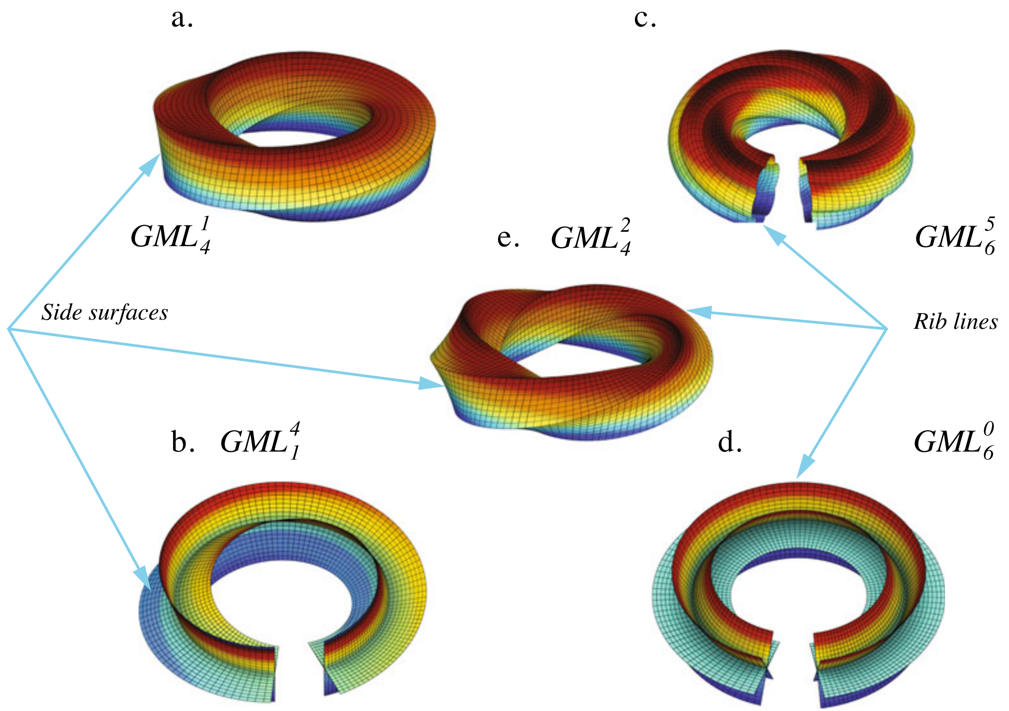


Fig. 2 Some examples of GML_m^n surfaces and bodies

μ (examples of rib lines: for Fig. 2a and b one torus line with $\mu = 1/4$; for Fig. 2c one torus line with $\mu = 5/6$; for Fig. 1e four and for Fig.2d seven circles i.e. for each—$\mu = 0$, for Fig.2f two torus lines with $\mu = 1/2$).

Definition 3. **A Side** of the $GML_m^n\{\mu\}$ body for each $GML_m^n\{\mu\}$, is a continuous closed surface, in which only the sides of the radial cross sections (plane m-symmetric figures) of this body or sides of prism PR_m are situated. These are the zones between the ribs.

To facilitate the calculation of long and complicated expressions, we introduce some notations and abbreviations:

$$\begin{gathered}
X \equiv X(\tau,\psi,\theta,t);\quad Y \equiv Y(\tau,\psi,\theta,t);\quad Z \equiv Z(\tau,\psi,\theta,t);\\
R \equiv R(\psi,\theta,t);\;\; R_\psi \equiv \frac{\partial R(\psi,\theta,t)}{\partial\psi};\;\; R_\theta \equiv \frac{\partial R(\psi,\theta,t)}{\partial\theta};\;\; Q \equiv Q(\theta,t);\\
p \equiv p(\tau,\psi,\theta,t);\;\; p_\tau \equiv \frac{\partial p(\tau,\psi,\theta,t)}{\partial\tau};\;\; p_\psi \equiv \frac{\partial p(\tau,\psi,\theta,t)}{\partial\psi};\;\; p_\theta \equiv \frac{\partial p(\tau,\psi,\theta,t)}{\partial\theta};\\
\overline{cos} \equiv \cos(\psi+\mu g(\theta));\quad \overline{sin} \equiv \sin(\psi+\mu g(\theta));\\
\underline{Cos} \equiv \cos(\theta+M(t));\quad \underline{Sin} \equiv \sin(\theta+M(t));\\
\frac{\partial\overline{cos}}{\partial\theta} \equiv \frac{\partial\cos(\psi+\mu g(\theta))}{\partial\theta} = -\mu g'(\theta)\sin(\psi+\mu g(\theta)) \equiv -\mu g'\cdot\overline{sin};\\
\frac{\partial\overline{sin}}{\partial\theta} \equiv \frac{\partial\sin(\psi+\mu g(\theta))}{\partial\theta} = \mu g'(\theta)\cos(\psi+\mu g(\theta)) \equiv \mu g'\cdot\overline{cos};
\end{gathered}\tag{7}$$

where g'—is the derivative of the function $g(\theta)$ with argument θ; in these abbreviations, representation (5) in static has the form

$$\begin{aligned}
X &= T_1 + [R + p\cdot\overline{cos}]\cdot\underline{Cos}\\
Y &= T_2 + [R + p\cdot\overline{cos}]\cdot\underline{Sin}\\
Z &= T_3 + Q + p\cdot\overline{sin},
\end{aligned}\tag{5*}$$

where $T_1 = T_1(t_0);\; T_2 = T_2(t_0);\; T_3 = T_3(t_0)$ are constants, defining the position in space.

• Some additional information about the classification of GRT_m^n and GML_m^n bodies are reported in [2–9].

III. Some Geometric Properties of “Semi-Regular” GRT_m^n or GML_m^n Bodies and Surfaces

In this part we study some geometric characteristic of a “Semi-Regular” static Generalized Möbius-Listing’s bodies GML_m^n. According to abbreviations (5*):

$$X_\tau \equiv \frac{\partial X}{\partial\tau} = p_\tau\cdot\overline{cos}\cdot\underline{Cos};\quad Y_\tau \equiv \frac{\partial Y}{\partial\tau} = p_\tau\cdot\overline{cos}\cdot\underline{Sin};\quad Z_\tau \equiv \frac{\partial Z}{\partial\tau} = p_\tau\cdot\overline{sin};\tag{8}$$

$$\begin{gathered}
X_\theta \equiv \frac{\partial X}{\partial \theta} = R_\theta \underline{Cos} - R\underline{Sin} + p_\theta \cdot \overline{cos} \cdot \underline{Cos} - \mu g' p \cdot \overline{sin} \cdot \underline{Cos} - p \cdot \overline{cos} \cdot \underline{Sin} \\
Y_\theta \equiv \frac{\partial Y}{\partial \theta} = R_\theta \underline{Sin} + R\underline{Cos} + p_\theta \cdot \overline{cos} \cdot \underline{Sin} - \mu g' p \cdot \overline{sin} \cdot \underline{Sin} + p \cdot \overline{cos} \cdot \underline{Cos} \\
Z_\theta \equiv \frac{\partial Z}{\partial \theta} = p_\theta \cdot \overline{sin} + \mu g' p \cdot \overline{cos}
\end{gathered} \tag{9}$$

$$\begin{gathered}
X_\psi \equiv \frac{\partial X}{\partial \psi} = R_\psi \underline{Cos} + p_\psi \cdot \overline{cos} \cdot \underline{Cos} - p \cdot \overline{sin} \cdot \underline{Cos} \\
Y_\psi \equiv \frac{\partial Y}{\partial \psi} = R_\psi \underline{Sin} + p_\psi \cdot \overline{cos} \cdot \underline{Sin} - p \cdot \overline{sin} \cdot \underline{Sin} \\
Z_\psi \equiv \frac{\partial Z}{\partial \psi} = p_\psi \cdot \overline{sin} + p \cdot \overline{cos}
\end{gathered} \tag{10}$$

Proposition 1. The Jacobi Matrix

$$J(\tau, \psi, \theta, t_0) = \begin{pmatrix} \frac{\partial X}{\partial \tau} & \frac{\partial Y}{\partial \tau} & \frac{\partial Z}{\partial \tau} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} & \frac{\partial Z}{\partial \theta} \\ \frac{\partial X}{\partial \psi} & \frac{\partial Y}{\partial \psi} & \frac{\partial Z}{\partial \psi} \end{pmatrix} \tag{11}$$

of all GTR^n_m bodies (according to representation (5) and (5*)) for all fixed values of time t_0 has determinant

$$det(J(\tau, \psi, \theta, t_0)) = p_\tau \cdot [R + p \cdot \overline{cos}] \cdot [p - R_\psi \cdot \overline{sin}] \tag{12}$$

Corollary 1. If $R_\psi = 0$ and $R > p$, i.e. 1. function $R(\psi, \theta)$ is independent of argument ψ; and 2. the large radius is greater than the radius of radial cross section of GRT^n_m bodies (both are completely natural conditions), then representation (5) is a one to one correspondence of the points of the PR_m prism and points of corresponding GRT^n_m body.

Proof. According to expression (12)

$$det((\tau, \psi, \theta, t_0)) = p \cdot p_\tau \cdot [R + p \cdot \overline{cos}] \neq 0$$

1. $p_\tau \neq 0$—natural condition (the opposite would mean that the function is independent of argument τ and this is impossible).

2. $p \neq 0$—otherwise this means the radial cross section is a point!

3. $R + p\overline{cos} > 0$ because $R > p > 0$.

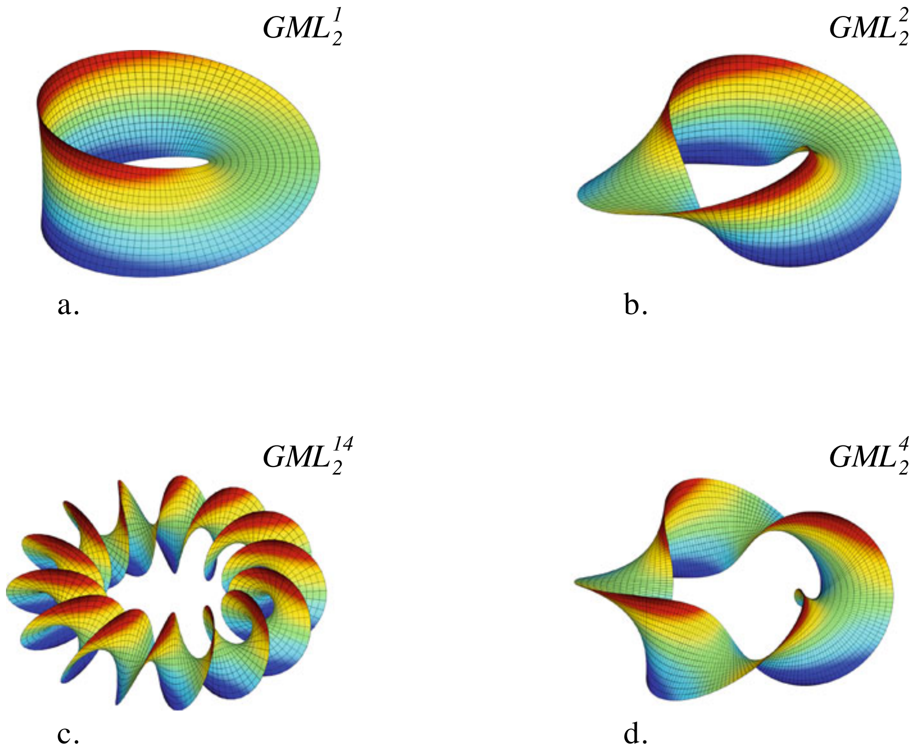


Fig. 3 Some examples of GML_2^n surfaces

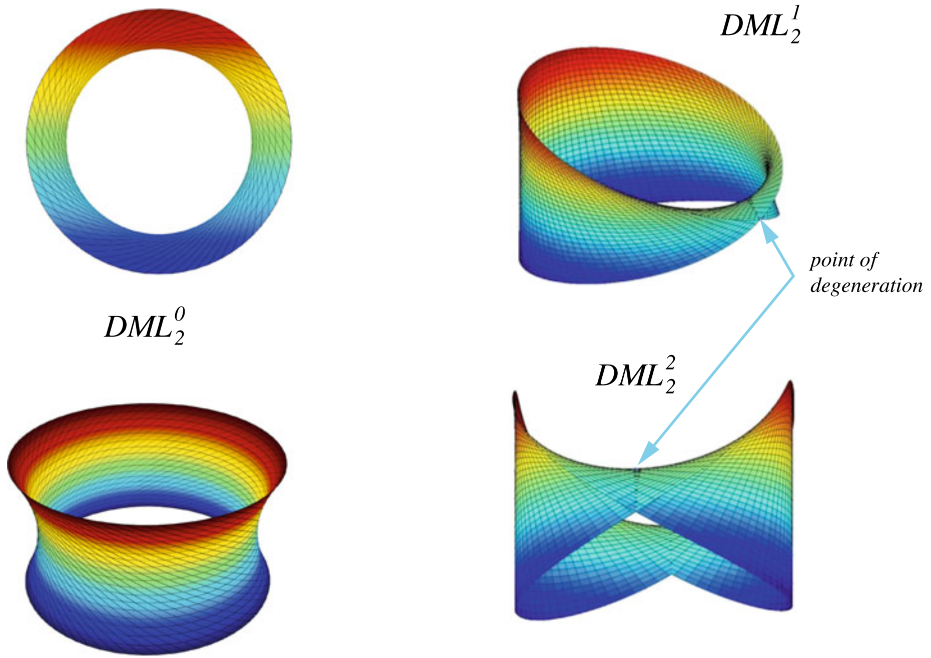


Fig. 4 Some examples of DML_2^n surfaces

Remark 1. **(A)** At any fixed moment t_0, this is a representation of a real three-dimensional body, which can be theoretically constructed from a prism PR_m by continuous deformation. This GRT_m^n or GML_m^n object (5) has no self-intersecting points.

(B) The representation (5) can be considered as a complex motion of a three-dimensional body or plane surface in time as a superposition of elementary displacements. (Example in representation (5), replace argument θ with argument t and the functions do not depend on argument θ).

(C) Roughly speaking the first movement is a twisting in a plane perpendicular to the main movement (the simplest example is the torsion of an aircraft propeller)!

(D) The GML_m^n surfaces and bodies are a very important subset of GRT_m^n bodies ($Q(\theta, t) = 0$ or $Q(\theta, t)$ - is a 2π-periodic sufficiently smooth function of argument θ).

Proposition 2. **1.** If $gcd(m, n) = k$, then a full external side of corresponding GML_m^n surface or body (Definition 3) (with radial cross section convex polygon) is a k—colored surface; i.e. it is possible to paint the surface of this figure in k different colors without taking away of the brush. It is prohibited to cross the rib of this figure [4–6].

2. If $gcd(m, n) = k$, then a full side of corresponding GML_m^n surface or body (with radial cross section simple star) is a $2k$—colored surface (see Remark 4 in [4, 5]);

Examples of different surfaces or bodies are shown in: Fig. 1e—4 colored; Fig. 2a, c and 3a are one colored, Fig. 2b, f, 3b, c and d are 2 colored; Fig. 2d—14 colored;

IV. Some Geometric Properties of "Regular" GML_2^n Surfaces

This is one of a simplest but most important subclass of GML surfaces, when the shape of the basic line does not depend on arguments ψ, θ and $g(\theta) \equiv \theta$. Some examples are given in Figure 3. They can be considered in two ways:

1. This is the static surface that is obtained from the rectangle PR_2 by twisting n times around the axis OO' of symmetry before gluing the ends (representation (13), when arguments ψ_0, t_0 are fixed). This means that the representation (5) has following simple form:

$$\begin{aligned} X(\tau, \psi_0, \theta, t_0) &= T_1(t_0) + [R + \tau \cdot \cos(\psi_0 + \frac{n}{2} \cdot \theta)] \cos(\theta + M(t_0)) \\ Y(\tau, \psi_0, \theta, t_0) &= T_2(t_0) + [R + \tau \cdot \cos(\psi_0 + \frac{n}{2} \cdot \theta)] \sin(\theta + M(t_0)) \\ Z(\tau, \psi_0, \theta, t_0) &= T_3(t_0) + \tau \cdot \sin(\psi_0 + \frac{n}{2} \cdot \theta) \end{aligned} \tag{13}$$

2. This is a trace (trajectory surface) that a segment leaves that revolves around a basic line perpendicular to it (particularly, representation (13*), when arguments ψ_0, θ_0 are fixed).

$$\begin{aligned} X(\tau, \psi_0, \theta_0, t) &= [R + \tau \cdot \cos(\psi_0 + \frac{n}{2} \cdot t] \cos(\theta_0 + t) \\ Y(\tau, \psi_0, \theta_0, t) &= [R + \tau \cdot \cos(\psi_0 + \frac{n}{2} \cdot t] \sin(\theta_0 + t)) \\ Z(\tau, \psi_0, \theta_0, t) &= \tau \cdot \sin(\psi_0 + \frac{n}{2} \cdot t) \end{aligned} \tag{13*}$$

Remark 2. **1.** According to Corollary 1, expressions: (12), (13) and $p = \tau$, $p_\tau = 1$,

$$det(J_{GML_2^n}(\tau, \psi_0, \theta, t_0)) = \tau \cdot [R + \tau \cdot \cos(\psi_0 + \frac{n}{2} \cdot \theta)]$$

and this expression never vanishes, since the condition is always $R > \tau > 0$ (because without this restriction it is impossible to make this object).

2. According to expression (13*) this trajectory may have some self-crossing points, because such a trajectory does not need such a restriction $R > \tau$.

3. According to Proposition 2 for GML_2^n surfaces:

a. If n is an even number, then each function X, Y, Z in the representations (13) or (13*) are 2π-periodic functions of the argument θ or t. (see e.g. Fig. 3b, c and d);

b. If n is a odd number, then each function X, Y, Z in the representation (13) is a 4π-periodic function satisfying the following properties (Möbius-property, see [6–8]) (see e.g. Fig. 3a).

$$X(\tau, \theta + 2\pi); Y(\tau, \theta + 2\pi); Z(\tau, \theta + 2\pi) = X(-\tau, \theta); Y(-\tau, \theta); Z(-\tau, \theta) \tag{14}$$

V. Some Examples and Geometric Properties of "Semi-Regular" DRT_m^n and DML_m^n Surfaces

These are the trajectories of bodies or surface which appear when:

1. DRT_m^n—a plane m-symmetrical figure (or PR_m-prism) makes n-turns around the baseline (only in the tangent plane of the virtual cylinder) after one complete round-trip of this curve around the axis OZ;

2. DML_m^n—a plane m-symmetrical figure makes n-turns around the baseline (only in the tangent plane of the virtual cylinder) before gluing.

We will call these geometric objects Degenerated Rotated and Twisted DRT_m^n and Degenerated Möbius-Listing's DML_m^n surfaces or bodies. The analytic representations of these motions according to abbreviations (7) have following form

$$\begin{aligned} X &= T_1 + R \cdot \underline{Cos} + p \cdot \overline{cos} \cdot \underline{Sin} \\ Y &= T_2 + R \cdot \underline{Sin} + p \cdot \overline{cos} \cdot \underline{Cos} \\ Z &= T_3 + Q + p \cdot \overline{sin}. \end{aligned} \tag{15}$$

The main difference between this representation (15) and (5) is not only in mathematical form, but also in the value of the determinant of the Jacobi matrix (in this case it is sometime reset to zero).

Proposition 3. The Jacobi matrix of all DTR^n_m bodies (according to representation (15)) for all fixed value of time t_0 have determinant

$$\begin{aligned} det(J_{DRT}(\tau,\psi,\theta,t_0)) &= p\cdot p_\tau\cdot[R_\theta-\mu g' R_\psi]\cdot[\underline{Sin^2}-\underline{Cos^2}]-p^2\cdot p_\tau\cdot\overline{cos} \\ &-R_\psi\cdot R\cdot p_\tau\cdot\overline{sin}+2p\cdot p_\tau\cdot\underline{Sin}\cdot\underline{Cos}\cdot[R+R_\psi\cdot\overline{sin}\cdot\overline{cos}] \qquad (16) \end{aligned}$$

Corollary 2. Even the simplest case determinant of Jacobi matrix for some values of the argument is zero! This is the point of degeneration on the surface.

$$det(J_{DRT}(\tau,\psi,\theta,t_0)) = p\cdot p_\tau\cdot[2R\cdot\underline{Sin}\cdot\underline{Cos}-p\cdot\overline{cos}].$$

Despite the conditions of the functions $R(\psi,\theta,t)$ and $p(\tau,\psi,\theta,t)$ there always remains the possibility that the determinants will be zero (Fig. 4).

Studying the geometric properties of these surfaces is certainly possible, but this can only be done if necessary, and then only in those sub-domains where there are no points of degenerations.

References

1. Tavkhelidze, I.: On the some properties of one class of geometrical figures. Bull. TICMI, Tbilisi **4**, 51–55 (2000)
2. Tavkhelidze, I., Ricci, P.E.: Classification of a wide set of geometric figures, surfaces and lines (trajectories). In: Rendiconti Accademia Nazionale delle Science detta dei XL, Memorie di Matematica e Applicazioni, 124^o, vol. XXX, fasc. 1, pp. 191–212 (2006)
3. Cassisa, C., Tavkhelidze, I.: About some geometric characteristic of the generalized Möbius-listing's surfaces. Georgian Electron. Sci. J.: Comput. Sci. Telecommun. **4**(21), 54–84 (2009)
4. Tavkhelidze, I.: About connection of the generalized Möbius-listing's surfaces with sets of ribbon knots and links. In: Ukrainian Mathematical Congress - 200, Section 2. Topology and Geometry, Proceedings of Institute of Mathematics Academy of Sciences of Ukraine, pp. 177–190 (2011). (in Ukrainian)
5. Tavkhelidze, I., Cassisa, C., Gielis, J., Ricci, P.E.: About "bulky" links, generated by generalized Möbius-listing's bodies GML^n_3. Rendiconti Lincei Mat. Appl. **24**, 11–38 (2013)
6. Tavkhelidze, I., Caratelli, D., Gielis, J., Ricci, P.E., Rogava, M., Transirico, M.: On a geometric model of bodies with "complex" configuration and some movements. In: Modeling in Mathematics. Atlantis Transactions in Geometry, vol. 2, pp. 129–159. Springer (2017). (Chapter 10)
7. Pinelas, S., Tavkhelidze, I.: Analytic representation of generalized Möbius-listing's bodies and classification of links appearing after their cut. In: Differential and Difference Equations with Applications (ICDDEA), Amadora, Portugal, June 2017. Springer Proceedings in Mathematics and Statistics, vol. 230, pp. 477–494 (2017)
8. Gielis, J., Tavkhelidze, I.: The general case of cutting of GML surfaces and bodies, pp. 1–75 (2019). https://arxiv.org/ftp/arxiv/papers/1904/1904.01414
9. Tavkhelidze, I., Gielis, J.: Structure of the d_m-knives and process of cutting of GML^n_m or GRT^n_m bodies. In: Reports of Enlarged Sessions of the Seminar of I. Vekua Institute of Applied Mathematics, vol. 33 (2019)
10. Gielis, J.: A generic geometric transformation that unifies a wide range of natural and abstract shapes. Am. J. Bot. **90**(3), 333–338 (2003)
11. Gray, A., Albena, E., Salamon, S.: Modern Differential Geometry of Curves and Surfaces with Mathematica, 3rd Edn. J. Capman and Hall/CRC

On Nonpower-Law Asymptotic Behavior of Blow-Up Solutions to Emden-Fowler Type Higher-Order Differential Equations

I. V. Astashova and M. Yu. Vasilev

Abstract For the equation

$$y^{(n)} = p_0 \mid y \mid^k \operatorname{sgn} y, \quad n \geq 12, \quad k > 1, \quad p_0 > 0, \tag{1}$$

the existence of positive solutions with nonpower-law asymptotic behavior is proved, namely

$$y(x) = (x^* - x)^{-\frac{n}{k-1}}\, h(\log(x^* - x)), \quad x \to x^* - 0, \tag{2}$$

where h is a positive periodic non-constant function on $\mathbb{R}$. To prove the existence, a useful modification of the Hopf bifurcation theorem is used.

1 Introduction

For the equation

$$y^{(n)} = p_0 \mid y \mid^k \operatorname{sgn} y, \quad n \geq 2, \quad k > 1, \quad p_0 > 0, \tag{3}$$

we study blow-up solutions, i.e. those with $\lim_{x \to x^* - 0} y(x) = \infty$. The origin of the considered problem is described in [1] (problem 16.4), and [2]. It was earlier proved for sufficiently large n (see [3]), for $n = 12$ (see [4]), for $n = 13, 14$ (see [5]), and for $n = 15$ (see [6]) that there exists $k = k(n) > 1$ such that Eq. (3) has a solution with nonpower-law asymptotic behavior, namely

I. V. Astashova (✉)
Lomonosov Moscow State University, Plekhanov Russian University of Economics, Moscow, Russian Federation
e-mail: ast.diffiety@gmail.com

M. Yu. Vasilev
Lomonosov Moscow State University, Moscow, Russian Federation
e-mail: vmuumv33@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_28

$$y(x) = (x^* - x)^{-\frac{n}{k-1}}\, h(\log(x^* - x)), \quad x \to x^* - 0, \tag{4}$$

where h is a positive periodic non-constant function on $\mathbb{R}$. Now we prove this result for arbitrary $n \geq 12$.

Note that it was also proved for $n = 2$ (see [1]) and for $n = 3, 4$ [7] that all blow-up solutions have power-law asymptotic behavior:

$$y(x) = C(x^* - x)^{-\alpha}\,(1 + o(1))\,, \quad x \to x^* - 0, \tag{5}$$

with

$$\alpha = \frac{n}{k-1}, \quad C = \left(\frac{\alpha(\alpha+1)\ldots(\alpha+n-1)}{p_0}\right)^{\frac{1}{k-1}}. \tag{6}$$

The existence of a solution satisfying (5) was proved for arbitrary $n \geq 2$. For $2 \leq n \leq 11$ an $(n-1)$-parametric family of such solutions to Eq. (3) was proved to exist for any x^* (see [7–9], Ch.**I**(5.1)). It was proved that for slightly superlinear equations of arbitrary order $n \geq 5$ all blow-up solutions have power-law asymptotic behavior (see [10, 11]), but for strongly superlinear equations of arbitrary order 12 to 203 (see [12]) a power-law asymptotic behavior is atypical (the Lebesgue measure of the set of initial data generating solutions with power-law asymptotic behavior is equal to zero).

2 Main Result

In this section, a result on the existence of solutions with nonpower-law asymptotic behavior is formulated for Eq. (3) with $n \geq 12$.

Theorem 1. *For any $n \geq 12$ there exists $k > 1$ such that Eq. (3) has a solution $y(x)$ with*

$$y^{(j)}(x) = (x^* - x)^{-\alpha-j}\, h_j(\,\log(x^* - x)\,),$$

$$j = 0, 1, \ldots, n-1,$$

where α is defined by (6) and h_j are periodic positive non-constant functions on $\mathbb{R}$.

3 Proof of the Main Result

To prove the main result of this article we transform Eq. (3) into a dynamical system and use a version of the Hopf Bifurcation theorem (see [13]).

3.1 Transformation of Equation (3)

Equation (3) can be transformed into a dynamical system (see [7] or [9], Ch.**I**(5.1)) by using the substitution

$$x^* - x = e^{-t}, \qquad y = (C + v)\, e^{\alpha t}, \tag{7}$$

where C and α are defined by (6). The derivatives $y^{(j)}$, $j = 0,\ 1, \dots, n-1$, become

$$e^{(\alpha+j)t} \cdot L_j(v,\ v', \dots, v^{(j)}),$$

where $v^{(j)} = \dfrac{d^j v}{dt^j}$, and L_j is a linear function with

$$L_j(0,\ 0,\ \dots, 0) = C\alpha(\alpha+1)\dots(\alpha+j-1) \neq 0$$

and its coefficient of $v^{(j)}$ equal to 1.

Thus (3) is transformed into

$$e^{(\alpha+n)t} \cdot L_n(v,\ v', \dots, v^{(n)}) = p_0\,(C+v)^k e^{\alpha k t} \tag{8}$$

and then into

$$v^{(n)} = p_0\,(C+v)^k - p_0\, C^k - \sum_{j=0}^{n-1} a_j v^{(j)}, \tag{9}$$

where $a_j,\ \ j = 1, ..., n,$ are the coefficients of $v^{(j)}$ in the linear function L_n and $(n-j)$-degree polynomial functions in α. Equation (9) can be written as

$$v^{(n)} = kC^{k-1} p_0 v - \sum_{j=0}^{n-1} a_j v^{(j)} + f(v), \tag{10}$$

where

$$f(v) = p_0\left((C+v)^k - C^k - kC^{k-1}v\right) = O(v^2),$$
$$f'(v) = O(v) \qquad \text{as} \quad v \to 0,$$

Suppose $V = (V_0, ..., V_{n-1})$ is the vector with coordinates $V_j = \ v^{(j)}$, $j = 0, \dots,$ $n-1$. Then Eq. (10) can be written as

$$\frac{dV}{dt} = AV + F(V), \tag{11}$$

where F is the vector function $F(V) = (0, ..., 0, F_{n-1}(V))$ with $F_{n-1}(V) = f(V_0)$ and A is a constant $n \times n$ matrix, namely

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \dots & \cdot \\ 0 & 0 & 0 & 0 & \dots & 1 \\ -\tilde{a}_0 & -a_1 & -a_2 & -a_3 & \dots & -a_{n-1} \end{pmatrix}$$

with

$$\begin{aligned} \tilde{a}_0 &= a_0 - kC^{k-1}p_0 = a_0 - k\alpha(\alpha+1)\dots(\alpha+n-1) \\ &= a_0 - (\alpha+1)\dots(\alpha+n-1)(\alpha+n) \end{aligned} \tag{12}$$

and eigenvalues satisfying the equation

$$\begin{aligned} 0 = \det(A - \lambda E) &= (-1)^{n+1}(-\tilde{a}_0 - a_1\lambda - \dots - a_{n-1}\lambda^{n-1} - \lambda^n) \\ &= (-1)^{n+1}\big((\alpha+1)(\alpha+2)\dots(\alpha+n) - (\lambda+\alpha)\dots(\lambda+\alpha+n-1)\big), \end{aligned} \tag{13}$$

which is equivalent to

$$\prod_{j=0}^{n-1}(\lambda + \alpha + j) = \prod_{j=0}^{n-1}(1 + \alpha + j). \tag{14}$$

3.2 Preliminary Results

Theorem 2 (The Hopf Bifurcation Theorem [14]**).** *Consider an α-parameterized dynamical system $\dot{x} = L_\alpha x + Q_\alpha(x)$ in a neighborhood of $0 \in \mathbb{R}^n$ with linear operators L_α and smooth enough functions $Q_\alpha(x) = O\left(|x|^2\right)$ as $x \to 0$. Let λ_α and $\bar{\lambda}_\alpha$ be simple complex conjugated eigenvalues of the operators L_α. Suppose $\mathrm{Re}\lambda_{\tilde{\alpha}} = \mathrm{Re}\bar{\lambda}_{\tilde{\alpha}} = 0$ for some $\tilde{\alpha}$ and the operator $L_{\tilde{\alpha}}$ has no other eigenvalues with zero real part.*

If $\mathrm{Re}\frac{d\lambda_\alpha}{d\alpha}(\tilde{\alpha}) \neq 0$, then there exist continuous mappings $\varepsilon \mapsto \alpha(\varepsilon) \in \mathbb{R}$, $\varepsilon \mapsto T(\varepsilon) \in \mathbb{R}$, and $\varepsilon \mapsto b(\varepsilon) \in \mathbb{R}^n$ defined in a neighborhood of 0 and such that $\alpha(0) = \tilde{\alpha}$, $T(0) = 2\pi/\mathrm{Im}\lambda_{\tilde{\alpha}}$, $b(0) = 0$, $b(\varepsilon) \neq 0$ for $\varepsilon \neq 0$, and the solutions to the problems

$$\dot{x} = L_{\alpha(\varepsilon)}x + Q_{\alpha(\varepsilon)}(x), \qquad x(0) = b(\varepsilon)$$

are $T(\varepsilon)$-periodic and non-constant.

Theorem 3 (Modification of the Hopf Theorem [13]**).** *Consider an α-parameterized dynamical system $\dot{x} = f(x, \alpha)$, where $f : \mathbb{R}^{n+1} \mapsto \mathbb{R}^n$ is a C^r function $(r \geq 3)$ such that $f(0, \alpha) = 0$ for all $\alpha \in \mathbb{R}$. Suppose the Jacobian matrix $D_x f(0, \tilde{\alpha}) \equiv A(\tilde{\alpha})$*

has $\pm i\beta$ *as simple eigenvalues for some* $\tilde{\alpha} \in \mathbb{R}$. *Let* v *and* w *be eigenvectors such that* $Av = \beta i v$, $A^* w = \beta i w$, *where* A^* *denotes the transpose conjugate matrix of the matrix* A. *Put* $\varphi \equiv \mathrm{Re}(e^{it} v)$, $\psi \equiv \mathrm{Re}(e^{it} w)$, $\Theta_j = \dfrac{1}{j!} \int_0^{2\pi} \left(\dfrac{\partial^j (f_x)}{\partial \alpha^j}(0, \tilde{\alpha})\varphi, \psi \right) dt$.

If $\Theta_c \neq 0$ *for some odd number* c, *then* $(0, \tilde{\alpha})$ *is a bifurcation point of periodic solutions of* $\dot{x} = f(x, \alpha)$. *More precisely, there exist continuous mappings* $\varepsilon \mapsto \alpha(\varepsilon) \in \mathbb{R}$, $\varepsilon \mapsto T(\varepsilon) \in \mathbb{R}$, *and* $\varepsilon \mapsto b(\varepsilon) \in \mathbb{R}^n$ *defined in a neighborhood of* 0 *and such that* $\alpha(0) = \tilde{\alpha}$, $T(0) = \frac{2\pi}{\beta}$, $b(0) = 0$, $b(\varepsilon) \neq 0$ *for* $\varepsilon \neq 0$, *and the solutions to the problems*

$$\dot{x} = f(x, \alpha(\varepsilon)), \qquad x(0) = b(\varepsilon)$$

are $T(\varepsilon)$*-periodic and non-constant.*

To apply the Hopf Bifurcation theorem we study Eq. (11) and the roots of the algebraic Eq. (14).

Lemma 1 *([5]). For any integer* $n \geq 12$ *there exist* $\alpha > 0$ *and* $q > 0$ *such that*

$$\prod_{j=0}^{n-1} (qi + \alpha + j) = \prod_{j=0}^{n-1} (1 + \alpha + j). \tag{15}$$

Lemma 2 *([5]). For any* $\alpha > 0$ *and any integer* $n > 1$ *all roots* $\lambda \in \mathbb{C}$ *to Eq. (14) are simple.*

3.3 *Proof of Theorem 1 for n* = 16

To apply the classical Hopf bifurcation theorem it remains to check the transversality condition $\mathrm{Re}\frac{d\lambda_\alpha}{d\alpha}(\tilde{\alpha}) \neq 0$. However, for this we have to prove a particular case of Lemma 1 (for $n = 16$) with the additional estimate $\alpha > 4$.

Consider the positive functions $\rho_n(\alpha)$ and $\sigma_n(\alpha)$ defined for all $\alpha > 0$ via the equations

$$\prod_{j=0}^{n-1} \left(\rho_n(\alpha)^2 + (\alpha + j)^2 \right) = \prod_{j=0}^{n-1} (1 + \alpha + j)^2 \tag{16}$$

and

$$\sum_{j=0}^{n-1} \arg \left(\sigma_n(\alpha) i + \alpha + j \right) = 2\pi \tag{17}$$

supposing $\arg z \in [0, 2\pi)$ for all $z \in \mathbb{C} \setminus \{0\}$.

Now, we prove the existence of $\alpha > 4$ such that $\rho_n(\alpha)$ and $\sigma_n(\alpha)$ are equal to the same value q, which makes the two sides of (15) to have equal modules squared and arguments.

Lemma 3. $\rho_n(\alpha) < \sigma_n(\alpha)$ *for sufficiently large* α.

Proof. Equation (16) defining the function $\rho_n(\alpha)$ may be written as

$$\prod_{j=0}^{n-1}\left(1+\frac{2j}{\alpha}+\frac{j^2}{\alpha^2}+\left(\frac{\rho_n(\alpha)}{\alpha}\right)^2\right)=\prod_{j=0}^{n-1}\left(1+\frac{j+1}{\alpha}\right)^2.$$

This shows that $\dfrac{\rho_n(\alpha)}{\alpha} \to 0$ as $\alpha \to +\infty$.

Equation (17) defining the function $\sigma_n(\alpha)$ may be written as

$$\sum_{j=0}^{n-1}\arctan\frac{\dfrac{\sigma_n(\alpha)}{\alpha}}{1+\dfrac{j}{\alpha}}=2\pi.$$

This shows that $\dfrac{\sigma_n(\alpha)}{\alpha} \to \tan\dfrac{2\pi}{n} > 0$ as $\alpha \to +\infty$. Thus, $\rho_n(\alpha) < \sigma_n(\alpha)$ for sufficiently large α.

Lemma 4. $\rho_{16}(4) > \sigma_{16}(4)$.

Proof. Calculations show that

$$\prod_{j=0}^{15}(\rho_{16}(4)^2+(4+j)^2)=\prod_{j=0}^{15}(5+j)^2=\frac{20!^2}{4!^2}>1.02\cdot 10^{34},$$

$$\prod_{j=0}^{15}((\sqrt{16.9})^2+4+j)^2)=\prod_{j=0}^{15}((16.9+(4+j)^2)<1.02\cdot 10^{34}.$$

Consequently, $\rho_{16}(4) > \sqrt{16.9}$. Now we prove that $\sigma_{16}(4) < \sqrt{16.9}$.

$$\sum_{j=0}^{15}\arg(i\sqrt{16.9}+4+j)=\sum_{j=4}^{19}\arctan\frac{\sqrt{16.9}}{j}=\sum_{j=4}^{11}\left(\arctan\frac{\sqrt{16.9}}{j}+\arctan\frac{\sqrt{16.9}}{23-j}\right)$$

$$=\left(\arctan\frac{\sqrt{16.9}}{4}+\arctan\frac{\sqrt{16.9}}{19}\right)+\left(\arctan\frac{\sqrt{16.9}}{5}+\arctan\frac{\sqrt{16.9}}{18}\right)$$

$$+\left(\arctan\frac{\sqrt{16.9}}{6}+\arctan\frac{\sqrt{16.9}}{17}\right)+\left(\arctan\frac{\sqrt{16.9}}{7}+\arctan\frac{\sqrt{16.9}}{16}\right)$$

$$+\left(\arctan\frac{\sqrt{16.9}}{8}+\arctan\frac{\sqrt{16.9}}{15}\right)+\left(\arctan\frac{\sqrt{16.9}}{9}+\arctan\frac{\sqrt{16.9}}{14}\right)$$

$$+\left(\arctan\frac{\sqrt{16.9}}{10}+\arctan\frac{\sqrt{16.9}}{13}\right)+\left(\arctan\frac{\sqrt{16.9}}{11}+\arctan\frac{\sqrt{16.9}}{12}\right)$$

$$=\arctan\frac{23\sqrt{16.9}}{59.1}+\arctan\frac{23\sqrt{16.9}}{73.1}+\arctan\frac{23\sqrt{16.9}}{85.1}+\arctan\frac{23\sqrt{16.9}}{95.1}$$

$$+\arctan\frac{23\sqrt{16.9}}{103.1}+\arctan\frac{23\sqrt{16.9}}{109.1}+\arctan\frac{23\sqrt{16.9}}{113.1}+\arctan\frac{23\sqrt{16.9}}{115.1}$$

$$=\left(\arctan\frac{23\sqrt{16.9}}{59.1}+\arctan\frac{23\sqrt{16.9}}{115.1}\right)+\left(\arctan\frac{23\sqrt{16.9}}{73.1}+\arctan\frac{23\sqrt{16.9}}{113.1}\right)$$

$$+\left(\arctan\frac{23\sqrt{16.9}}{85.1}+\arctan\frac{23\sqrt{16.9}}{109.1}\right)+\left(\arctan\frac{23\sqrt{16.9}}{95.1}+\arctan\frac{23\sqrt{16.9}}{103.1}\right)$$

$$>(\pi-\arctan 8)+(\pi-\arctan 27)+\arctan 53+\arctan 27$$

$$>\ 2\pi-\arctan 8+\arctan 53>2\pi.$$

Lemma 4 is proved.

It follows from the previous lemmas that for $n=16$ there exists $\alpha>4$ such that $\rho_n(\alpha)=\sigma_n(\alpha)$. In other words, there exists $\alpha>4$ and $q>0$ satisfying (15).

Lemma 5. *If $\alpha>4$ and $q>0$ satisfy the polynomial Eq.* (15) *with $n=16$, then $q^2<3\alpha+5$.*

Proof. Suppose that $q^2 \geq 3\alpha + 5$. Then by the substitution $\alpha = x + 4$, $x > 0$ we obtain

$$0 = \prod_{j=0}^{15}(q^2 + (\alpha + j)^2) - \prod_{j=0}^{15}(1 + \alpha + j)^2 \geq \prod_{j=0}^{15}(3\alpha + 5 + (\alpha + j)^2) - \prod_{j=0}^{15}(1 + \alpha + j)^2$$

$$= \prod_{j=0}^{15}(3(x + 4) + 5 + (x + 4 + j)^2) - \prod_{j=0}^{15}(5 + x + j)^2 = 16x^{31} + 6008x^{30} + 1083320x^{29}$$

$$\begin{gathered}
+ 124909060x^{28} + 10346958048x^{27} + 655873445556x^{26} + 33088544670480x^{25} \\
+ 1364273934048714x^{24} + 46845738509472552x^{23} + 1358250645535902456x^{22} \\
+ 33597563403075497280x^{21} + 714518112665170116810x^{20} \\
+13139980014580765008816x^{19} + 209824392142851028227096x^{18} \\
+ 2917526044055626441371336x^{17} + 35381047693929819567554847x^{16} \\
+374406309656399083867870608x^{15} + 3455282329668110634135853068x^{14} \\
+ 27762847831737562495107829808x^{13} + 193660610621079373247365814128x^{12} \\
+ 1167859506965501209338409024504x^{11} + 6053414704888747912009095383744x^{10} \\
+ 26761570080450447023578667222592x^{9} + 99878302212425543314300198979457x^{8} \\
+ 310418927289662193023860542327984x^{7} + 788696073282846162775363136169840x^{6} \\
+1596416517107600883219437082580800x^{5} + 2478942035386106131273894570306656x^{4} \\
+ 2782844558929121673466794032264448x^{3} + 2032786024054120635868885845636864x^{2} \\
+ 764150278182231899482437402897408x + 521010595519466252089690097256 96,
\end{gathered}$$

which is positive for any $x > 0$. This contradiction to (15) yields $q^2 < 3\alpha + 5$.

Lemma 5 is proved.

The condition $\operatorname{Re}\frac{d\lambda_\alpha}{d\alpha}(\tilde{\alpha}) \neq 0$ needed for the Hopf theorem, expressed explicitly by means of the implicit function theorem, looks like

$$\left[\sum_{j=0}^{n-1}\frac{\alpha + j}{q^2 + (\alpha + j)^2}\right]^2 + \left[\sum_{j=0}^{n-1}\frac{q}{q^2 + (\alpha + j)^2}\right]^2 \neq \sum_{j=0}^{n-1}\frac{\alpha + j}{q^2 + (\alpha + j)^2}\sum_{j=0}^{n-1}\frac{1}{1 + \alpha + j}.$$

Lemma 6. *If* $n = 16$, $0 < q^2 < 3\alpha + 5$, *then*

$$\left[\sum_{j=0}^{n-1} \frac{\alpha + j}{q^2 + (\alpha + j)^2}\right]^2 + \left[\sum_{j=0}^{n-1} \frac{q}{q^2 + (\alpha + j)^2}\right]^2 \\ > \sum_{j=0}^{n-1} \frac{\alpha + j}{q^2 + (\alpha + j)^2} \sum_{j=0}^{n-1} \frac{1}{1 + \alpha + j}. \tag{18}$$

Proof. Hereafter all sums and products with no limits indicated are over $j = 0, 1, \ldots, n-1$.

Multiplying inequality (18) by $U_* = \prod(1 + \alpha + j)$ and then twice by $V_* = \prod\left[q^2 + (\alpha + j)^2\right]$, we obtain the following equivalent inequality provided $\alpha > 0$:

$$U_*\left[\left(\sum(\alpha + j)V_j\right)^2 + q^2\left(\sum V_j\right)^2\right] > V_* \sum(\alpha + j)V_j \sum U_j \tag{19}$$

with the polynomials $U_j = \dfrac{U_*}{1 + \alpha + j}$ and $V_j = \dfrac{V_*}{q^2 + (\alpha + j)^2}$.

Put $q^2 = \dfrac{3\alpha + 5}{1 + w}$, $w > 0$. Substituting this into inequality (19) and multiplying the result by $(1 + w)^{2n-1}$ we obtain another equivalent one:

$$U_*\left[(1 + w)\left(\sum(\alpha + j)P_j\right)^2 + (3\alpha + 5)\left(\sum P_j\right)^2\right] \\ > P_* \cdot \sum(\alpha + j)P_j \cdot \sum U_j \tag{20}$$

with $P_* = \prod\left[3\alpha + 5 + (1 + w)(\alpha + j)^2\right]$ and $P_j = \dfrac{P_*}{3\alpha + 5 + (1 + w)(\alpha + j)^2}$.

Both sides of inequality (20) are polynomials of α and w with non-negative integer coefficients. So, they can be computed exactly, with no rounding. This rather cumbersome computation gives the following result for the difference of the left- and right-hand sides of (20) expressed as

$$U_*\left[(1 + w)\left(\sum(\alpha + j)P_j\right)^2 + (3\alpha + 5)\left(\sum P_j\right)^2\right] \\ - P_* \sum(\alpha + j)P_j \sum U_j = \sum_{j=0}^{5n-2} \Delta_j \alpha^j \tag{21}$$

with polynomials $\Delta_j \in \mathbb{R}[w]$. Straightforward though very cumbersome calculations show that $\Delta_{5n-2} = 0$, and all other Δ_j in (21) are polynomials with positive coefficients if $n = 16$.

This completes the proof of Lemma 6.

Now the Hopf bifurcation theorem and the lemmas proved provide, for $n = 16$, the existence of a family $\alpha_\varepsilon > 0$ such that Eq. (14) with $\alpha = \alpha_0$ has imaginary roots $\lambda = \pm qi$ and, for sufficiently small ε, system (11) with $\alpha = \alpha_\varepsilon$ has a periodic solution $V_\varepsilon(t)$ with period $T_\varepsilon \to T = \frac{2\pi}{q}$ as $\varepsilon \to 0$. In particular, the coordinate $V_{\varepsilon,0}(t) = v(t)$ of the vector $V_\varepsilon(t)$ is also a periodic function with the same period. Then, taking into account (7), we obtain

$$y(x) = \big(C + v(-\log(x^* - x))\big)\,(x^* - x)^{-\alpha}.$$

Put $h(s) = C + v(-s)$, which is a non-constant continuous periodic and positive for sufficiently small ε function and obtain the required equality

$$y(x) = (x^* - x)^{-\alpha}\, h(\,\log(x^* - x)\,).$$

In the similar way we obtain the related expressions for $y^{(j)}(x)$, $j = 0, \ldots, n-\ 1$.

Theorem 1 for $n = 16$ is proved.

3.4 *Proof of Theorem 1 in General Case*

We can obtain some useful formulas

$$\tilde{a}_0 = \alpha(\alpha+1)\ldots(\alpha+n-1) - (\alpha+1)\ldots(\alpha+n) = -n(\alpha+1)\ldots(\alpha+n-1), \quad (22)$$

$$\frac{d^{n-1}(-\tilde{a}_0)}{d\alpha^{n-1}} = n!, \quad \frac{d^{n-1}(-a_1)}{d\alpha^{n-1}} = -n!, \quad (23)$$

$$\frac{d^{n-2}(-\tilde{a}_0)}{d\alpha^{n-2}} = n\left((n-1)!\alpha + (n-2)!\frac{n(n-1)}{2}\right) = \frac{(2\alpha+1)n!}{2}, \quad (24)$$

$$\frac{d^{n-1}(-a_2)}{d\alpha^{n-1}} = 0, \quad \frac{d^{n-2}(-a_2)}{d\alpha^{n-2}} = -(n-2)!\frac{n(n-1)}{2} = -\frac{n!}{2}. \quad (25)$$

By using (13), we can prove for $n,\ \alpha,\ q$ from Lemma 1 that the vector

$$v = (1, qi, -q^2, -q^3 i, q^4, \ldots)$$

is an eigenvector of the matrix A corresponding to the eigenvalue qi. Consider also an eigenvector w of the matrix A^* corresponding to the eigenvalue qi, assuming its last coordinate to equal 1: $w = (\ldots\ldots, 1)$. Then

$$\varphi = \mathrm{Re}(e^{it}v) = (\cos t, -q\sin t, -q^2\cos t, q^3\sin t, q^4\cos t, \ldots),$$

$$\psi = \mathrm{Re}(e^{it}w) = (\ldots\ldots, \cos t).$$

Using formulas (23)–(25), we obtain

$$\Theta_{n-1} = \frac{1}{(n-1)!}\int_0^{2\pi} \left(\begin{pmatrix} 0 & 0 & 0 \ldots 0 \\ \vdots & \vdots & \vdots \; \vdots \; \vdots \\ 0 & 0 & 0 \ldots 0 \\ n! & -n! & 0 \ldots 0 \end{pmatrix} \begin{pmatrix} \cos t \\ -q \sin t \\ \vdots \\ \vdots \end{pmatrix}, \begin{pmatrix} \vdots \\ \vdots \\ \vdots \\ \cos t \end{pmatrix} \right) dt$$

$$= \frac{1}{(n-1)!}\int_0^{2\pi} n!\,(\cos^2 t + q \sin t \cos t)\,dt = \pi n \neq 0,$$

$$\Theta_{n-2} = \frac{1}{(n-2)!}\int_0^{2\pi} \left(\begin{pmatrix} 0 & 0 & 0 & 0 \ldots 0 \\ \vdots & \vdots & \vdots & \vdots \; \vdots \; \vdots \\ 0 & 0 & 0 & 0 \ldots 0 \\ \frac{(2\alpha+1)n!}{2} & \frac{d^{n-2}(-a_1)}{d\alpha^{n-2}} & -\frac{n!}{2} & 0 \ldots 0 \end{pmatrix} \begin{pmatrix} \cos t \\ -q \sin t \\ -q^2 \cos t \\ \vdots \end{pmatrix}, \begin{pmatrix} \vdots \\ \vdots \\ \vdots \\ \cos t \end{pmatrix} \right) dt$$

$$= \frac{\pi}{(n-2)!}\left(\frac{(2\alpha+1)n!}{2} + \frac{q^2 n!}{2} \right) = \frac{\pi n(n-1)}{2}(2\alpha + 1 + q^2) > 0.$$

So, for any $n \geq 12$ we have $\Theta_{n-1} > 0$ and $\Theta_{n-2} > 0$ if $\alpha > 0$, and one of the numbers $n-1$ or $n-2$ is odd. Consequently, all conditions of Theorem 3 due to the lemmas are fulfilled. Therefore, for $n \geq 12$ the existence of a family $\alpha_\varepsilon > 0$ was proved such that Eq. (14) with $\alpha_0 = \tilde{\alpha}$ has imaginary roots $\lambda = \pm qi$ and for sufficiently small ε system (11) with $\alpha = \alpha_\varepsilon$ has a periodic solution $V_\varepsilon(t)$ with period $T_\varepsilon \to T = \frac{2\pi}{q}$ as $\varepsilon \to 0$. This, as well as in previous section, completes the proof of Theorem 1, for arbitrary $n \geq 12$ now.

References

1. Kiguradze, I.T., Chanturia, T.A.: Asymptotic Properties of Solutions of Nonautonomous Ordinary Differential Equations. Kluwer Academic Publishers, Boston (1993)
2. Astashova, I.V.: On Asymptotic Behavior of Blow-Up Solutions to Higher-Order Differential Equations with General Nonlinearity, Differential and Difference Equations with Applications, pp 1–13. Springer (2018)
3. Kozlov, V.A.: On Kneser solutions of higher order nonlinear ordinary differential equations. Ark. Mat. **37**(2), 305–322 (1999)
4. Astashova, I.V., Vyun, S.A.: On positive solutions with non-power asymptotic behavior to Emden-Fowler type twelfth order differential equation. Diff. Eq. **48**(11), 1568–1569 (2012). (Russian)
5. Astashova, I.V.: On power and non-power asymptotic behavior of positive solutions to Emden-Fowler type higher-order equations. Adv. Diff. Eq. J. **2013**, 220 (2013)

6. Vasilev, M.Yu.: On positive solutions with nonpower-law behavior to Emden–Fowler 15th-order equations. In: Proceedings of International Youth Scientific Forum “Lomonosov-2018” ser. Electronic resource (DVD-ROM). Lomomosov MSU (2018). ISBN 978-5-317-05800-5
7. Astashova, I.V.: Asymptotic behavior of solutions of certain nonlinear differential equations. In: Reports of extended session of a seminar of the I. N. Vekua Institute of Applied Mathematics, Tbilisi, vol. 1, no. 3, pp. 9–11 (1985). (Russian)
8. Astashova, I.V.: Application of dynamical systems to the study of asymptotic properties of solutions to nonlinear higher-order differential equations. J. Math. Sci. **126**(5), 1361–1391 (2005)
9. Astashova, I.V.: Qualitative properties of solutions to quasilinear ordinary differential equations. In: Astashova, I.V (ed.) Qualitative Properties of Solutions to Differential Equations and Related Topics of Spectral Analysis, Scientific Edition, pp. 22–290. UNITY-DANA, Moscow (2012). (Russian)
10. Astashova, I.V.: On Kiguradze’s problem on power-law asymptotic behavior of blow-up solutions to Emden-Fowler type differential equations. Georgian Math. J. **24**(2), 185–191 (2017)
11. Astashova, I.V.: On asymptotic behavior of blow-up solutions to higher-order differential equations with general nonlinearity. In: Pinelas, S., Caraballo, T., Kloeden, P., Graef, J. (eds) Differential and Difference Equations with Applications, ICDDEA 2017. Springer Proceedings in Mathematics & Statistics, vol. 230, pp. 1–12. Springer Cham (2018)
12. Astashova, I.V.: Asymptotic behavior of singular solutions of Emden-Fowler type equations. Diff. Eq. **55**(5), 581–590 (2019)
13. Tan, N.X., Tu, P.N.V.: Some new Hopf Bifurcation Theorems at simple eigenvalues. Appl. Anal. **53**(3–4), 197–220 (1994)
14. Marsden, J.E., McCracken, M.: The Hopf Bifurcation and its Applications. Springer, New York (1976)

Third Order Iterative Method for Nonlinear Difference Schemes

Irina Iumanova and Svyatoslav Solodushkin

Abstract A partial differential equation with fractional Riesz derivative and nonlinearity in differentiation operator is studied. We considered an implicit method which is a fractional analogue of Crank-Nicolson method and, therefore, implies the necessity of iterative solving of non-linear system on each time layer. To accelerate the convergence we elaborate a two stage iterative method which does not use derivatives and could be considered as an analog of Stefensen's method. The theorem of third order convergence is proved. Results of numerical examples coincides with theoretical ones.

1 Introduction

Over the past decades it has been recognized that fractional partial differential equations are convenient mathematical tool for describing some phenomena in numerous fields such as viscoelasticity, control systems, population dynamics, financial problems and physics, see [4, 5, 7, 12] and loads of references therein. Due to the fact that many natural processes are non-linear it is necessary to study fractional differential equations in partial derivatives with non-linearity in differentiation operators. From the mathematical point of view these equations provide an example of extremely complex, interesting and little-studied object. Analytical solution of these equations could be found in exceptional cases only, and therefore the elaboration of appropriate numerical methods and acceleration of their convergence is a relevant problem.

Numerical methods for partial fractional differential equations with non-linearity in heterogeneous function, but not in differential operators, have been already built and deeply studied in many works [10].

I. Iumanova (✉)
Ural Federal University, Yekaterinburg, Russia
e-mail: irina.iumanova@urfu.ru

S. Solodushkin
N.N. Krasovskii Institute of Mathematics and Mechanics, Yekaterinburg, Russia
e-mail: s.i.solodushkins@urfu.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_29

In [16] authors focused on the study of the Crank–Nicolson scheme for the Riesz space fractional-order parabolic type sine-Gordon equation. The existence, uniqueness, stability, and convergence were proved.

Implicit difference schemes for fractional partial differential equations with time delay were constructed in [8, 9]. The authors used shifted Grünwald–Letnikov formulas for the approximation of fractional derivatives with respect to spatial variables and the L1-algorithm for the approximation of fractional derivatives in time.

A discrete monotone iterative method for space-fractional non-linear diffusion-reaction equation was reported in [5]. The authors proposed a Crank–Nicolson discretization of a reaction-diffusion system with fractional spatial derivative of the Riesz type. The finite-difference scheme was based on the use of fractional order centered differences, and it was solved by a monotone iterative technique. As an application, the particular case of the space-fractional Fisher's equation is theoretically analyzed in full detail. In that case, the monotone iterative method guarantees the preservation of the positivity and the boundedness of the numerical approximations.

On the other hand numerical methods for partial fractional differential equations with non-linearity in differentiation operators have not been studied yet. In [12], as in most similar works, numerical methods are not considered, but attempts are made to find the exact solution in the form of series.

For integer order partial differential equations with non-linearity in differential operators an implicit difference scheme was constructed in [11]. Then the non-linear difference scheme was solved by the Newton method and convergence was demonstrated in numerical experiments, but no proves were given.

The development of difference schemes for partial fractional differential equations with non-linearity in differentiation operators has some issues. Explicit schemes as it was demonstrated in numerical experiments are unstable. On the other hand direct application of the implicit scheme leads to the necessity to solve non-linear systems of high dimensional on each time layer. For integer order derivatives these systems are three-diagonal or at least band, and special iterative methods could be applied. For Riesz derivatives, which are nonlocal, these non-linear systems are fully-filled, and complexity of their solving increase dramatically. This is why the elaboration of an iterative method which requires less number of calculation to achieve the necessary accuracy seems very tempting.

Iterative methods are the main to solve non-linear problems [2, 3, 13]. Despite the fact that tremendous efforts were made to improve the computational efficiency of iterative methods, in particular to increase the convergence rate, there is a lack of theoretical investigation in multidimensional case. Namely, usually authors offer a new method, prove the theorem of convergence in Banach space and after that either estimate the order in series of numerical experiments or prove the convergence with order in $\mathbb{R}^1$ case only.

In [1] the Newton-type method was proposed and divided differences instead of Frechet derivatives were used. Theorems of semilocal convergence in a Banach space were proved, however the order of convergence was not proved.

We consider an initial boundary value problem

$$\frac{\partial \omega(u(x,t))}{\partial t} = \frac{\partial^{\alpha} u(x,t)}{\partial |x|^{\alpha}} + f(x,t), \tag{1}$$

where t and x are independent variables, $0 \leqslant t \leqslant T$, $0 \leqslant x \leqslant X$, $u(x,t)$ is an unknown function to be found and ω is a given function which properties are described below, the order α of fractional Riesz derivative is assumed to be in range $1 < \alpha \leqslant 2$. The Riesz fractional derivative is defined as follow

$$\frac{\partial^{\alpha} u(x,t)}{\partial |x|^{\alpha}} = \frac{1}{2\cos(\frac{\pi\alpha}{2})\Gamma(2-\alpha)} \frac{d^2}{dx^2} \int_{-\infty}^{+\infty} \frac{u(\xi,t)}{|x-\xi|^{\alpha-1}} d\xi. \tag{2}$$

Initial and boundary conditions are set as follow

$$u(x,0) = \varphi(x),\ 0 \leqslant x \leqslant X, \tag{3}$$

$$u(0,t) = \mu_0(t),\ u(X,t) = \mu_1(t),\ 0 \leqslant t \leqslant T. \tag{4}$$

It is supposed that $u(x,t) = 0$ for $x < 0$ and $x > X$.

We assume that the problem (1)–(4) has a unique solution, understood in the classical sense, and this solution sufficiently smooth. We also assume that ω is twice continuously differentiable in its domain and its first derivative is uniformly greater than zero in the neighbourhood of solution u

$$0 < \hat{\omega} \leq \omega'(u). \tag{5}$$

These conditionals are essential to ensure the convergence of the difference scheme, and details could be found in [6], where the convergence of a non-linear difference scheme with usage of Newton method on time layers was shown for initial boundary value problem with integer derivatives. For Riesz fractional derivatives the technique of proof is the same.

2 Implicit Difference Scheme

We consider an equidistant partition of $[0, X]$ into parts with step size $h = X/N$ and define the grid $x_i = ih$, $i = 0, \ldots, N$. We also split the time interval $[0, T]$ into M parts with step size $\Delta = T/M$ and define the grid $t_j = j\Delta, j = 0, \ldots, M$.

Denote by u_j^i the approximation of the function value $u(x_i, t_j)$, $i = 0, 1, \ldots N$, $j = 0, \ldots M$, at the respective node.

To approximate the Riesz fractional derivative in the internal grid nodes we use formula [15]

$$\frac{\partial^\alpha u(x_i,t_j)}{\partial|x|^\alpha} = -\frac{1}{h^\alpha}\sum_{s=-i}^{N-i} g_{\alpha,s}u_j^{i+s} + O(h^2), \quad 1 \le i \le N-1, \tag{6}$$

where $\{g_{\alpha,s}\}_{s=-\infty}^{+\infty}$ be a sequence defined by

$$g_{\alpha,s} = \frac{(-1)^s\Gamma(\alpha+1)}{\Gamma(\frac{\alpha}{2}-s+1)\Gamma(\frac{\alpha}{2}+s+1)}.$$

Let us consider a non-linear difference scheme, $j = 0, 1, \ldots, M-1$,

$$\begin{aligned}\frac{\omega(u_{j+1}^i)-\omega(u_j^i)}{\Delta} = \frac{-1}{2h^\alpha}\left(\sum_{s=-i}^{N-i} g_{\alpha,s}u_{j+1}^{i+s} + \sum_{s=-i}^{N-i} g_{\alpha,s}u_j^{i+s})\right) + f_{j+1/2}^i,& \\ \text{for } i = 1,\ldots,N-1,& \\ \text{and } u_{j+1}^0 = \mu_0(t_{j+1}),\ u_{j+1}^N = \mu_1(t_{j+1}),&\end{aligned} \tag{7}$$

with initial conditions $u_j^i = \varphi(x_i),\ i = 0, 1\ldots, N$. To make notation shorter $f_{j+1/2}^i$ denotes $f(x_i, t_j + \Delta/2)$.

To find a solution $u_{j+1} = (u_{j+1}^1, u_{j+1}^2, \ldots, u_{j+1}^{N-1})$ on each next time layer it is necessary to solve a corresponding non-linear system, which can be written in the form $F(u_{j+1}) = 0,\ F : \mathbb{R}^{N-1} \to \mathbb{R}^{N-1}$. To deal with it we build a two-step iterative method, and all necessary definitions are given below.

Definition 1 ([14]). Let F be a continuous non-linear mapping from a Banach space X to Y. Linear operator $F(u_1, u_2)$ is called the first divided difference of operator F if the following requirements hold:

1. for each fixed $u_1, u_2 \in X$ operator $F(u_1, u_2)$ is such that $F(u_1, u_2)(u_1 - u_2) = F(u_1) - F(u_2)$;
2. if the Frechet derivative $F'(u)$ exists then $F(u, u) = F'(u)$.

Firstly it should be mentioned that Definition (1) is given for a function, but not for an equation. Secondly it specifies the first divided difference in a not unique way. To give an example of one possible constructivization of Definition (1) let us consider the simplest multidimensional case $X = Y = \mathbb{R}^2$, a function $F(u) : \mathbb{R}^2 \to \mathbb{R}^2$ and two points $u_1 = (u_1^1, u_1^2)$, $u_2 = (u_2^1, u_2^2)$, so

$$F(u_1) = \begin{pmatrix} f_1(u_1^1, u_1^2) \\ f_2(u_1^1, u_1^2)\end{pmatrix}, \qquad F(u_2) = \begin{pmatrix} f_1(u_2^1, u_2^2) \\ f_2(u_2^1, u_2^2)\end{pmatrix}.$$

The first divided difference could be define as follow

$$F\left(u_1,\ u_2\right)=\begin{pmatrix}\dfrac{f_1(u_1^1,\ u_1^2)-f_1(u_2^1,\ u_1^2)}{u_1^1-u_2^1} & \dfrac{f_1(u_2^1,\ u_1^2)-f_1(u_2^1,\ u_2^2)}{u_1^2-u_2^2}\\ \dfrac{f_2(u_1^1,\ u_1^2)-f_2(u_2^1,\ u_1^2)}{u_1^1-u_2^1} & \dfrac{f_2(u_2^1,\ u_1^2)-f_2(u_2^1,\ u_2^2)}{u_1^2-u_2^2}\end{pmatrix}. \tag{8}$$

Now let us consider the equation $F(u)=0$ in a form of fixed point problem $\Phi(u)=u$, where $\Phi(u)=u-\lambda F(u)$ and λ is parameter. For the simplest 2-dimensional case, i.e. $u=(u^1,u^2)$, these equations could be respectively represented in coordinate-wise forms

$$\begin{cases}f_1(u^1,\ u^2)=0,\\ f_2(u^1,\ u^2)=0,\end{cases}\qquad \begin{cases}\varphi_1(u^1,\ u^2)=u^1,\\ \varphi_2(u^1,\ u^2)=u^2,\end{cases}$$

where

$$\varphi_1(u^1,\ u^2)=u^1-\lambda f_1(u^1,\ u^2),\quad \varphi_2(u^1,\ u^2)=u^2-\lambda f_2(u^1,\ u^2).$$

In term of Φ the definition of the first divided difference has the following form

$$\Phi\left(u_1,\ u_2\right)=\begin{pmatrix}\dfrac{\varphi_1(u_1^1,\ u_1^2)-\varphi_1(u_2^1,\ u_1^2)}{u_1^1-u_2^1} & \dfrac{\varphi_1(u_2^1,\ u_1^2)-\varphi_1(u_2^1,\ u_2^2)}{u_1^2-u_2^2}\\ \dfrac{\varphi_2(u_1^1,\ u_1^2)-\varphi_2(u_2^1,\ u_1^2)}{u_1^1-u_2^1} & \dfrac{\varphi_2(u_2^1,\ u_1^2)-\varphi_2(u_2^1,\ u_2^2)}{u_1^2-u_2^2}\end{pmatrix}. \tag{9}$$

To find u_{j+1} in difference scheme (7) we consider it in an abstract form $F(u_{j+1})=0,\ F:\mathbb{R}^{N-1}\to\mathbb{R}^{N-1}$. We also consider an auxiliary function: $\Phi(u_{j+1})=u_{j+1}-\lambda F(u_{j+1})$.

Let us propose a two-step iterative method

$$u_{j+1}^{(k+1)}=\widetilde{u}_{j+1}^{(k)}-\mu\left[F\left(u_{j+1}^{(k)},\Phi\left(u_{j+1}^{(k)}\right)\right)\right]^{-1}F\left(\widetilde{u}_{j+1}^{(k)}\right), \tag{10}$$

$$\widetilde{u}_{j+1}^{(k)}=u_{j+1}^{(k)}-\left[F\left(u_{j+1}^{(k)},\Phi\left(u_{j+1}^{(k)}\right)\right)\right]^{-1}F\left(u_{j+1}^{(k)}\right), \tag{11}$$

where $u_{j+1}^{(k)}$ is the k-th iteration of a solution u_{j+1}, and $\widetilde{u}_{j+1}^{(k)}$ is an auxiliary sequence, $\lambda\in(0,\ 1],\ \mu\in(0,\ 1],\ k=0,1,\ldots,K$. Without loss of generality and to simplify the narration we assume that the same number K of iterations are performed on each time layer. As initial approximation in (11) we take a value from the previous time layer $u_{j+1}^{(0)}=u_j^{(K)}$. Note that $F\left(u',u''\right)=\frac{1}{\lambda}\left(E-\Phi\left(u',u''\right)\right)$.

For initial boundary value problem with integer derivatives and non-linearity in differential operators the convergence of a non-linear difference scheme with usage of Newton method on time layers was shown in [6]. For Riesz fractional derivatives the technique of proof is the same. In the next section we show that method (10)–(11) converges with third order and its computational efficiency is greater then in Newton method.

3 Convergence of Two-Step Iterative Method

We consider an equation $F(u) = 0,\ F : \mathbb{R}^{N-1} \to \mathbb{R}^{N-1}$. Let $F(u)$ be a continues, the operator of the first divided difference is invertible, i.e. in the domain of our interest there exist $[F(u', u'')]^{-1} = [E - \Phi(u', u'')]^{-1}$. We consider method (10)–(11) and to simplify notation we will not use subscript $j + 1$; without loss of generality we also take $\mu = \lambda = 1$, see remark below.

Theorem 1. *Let the following conditions hold:*

1. $\left\|F(u^{(0)})\right\| = \left\|u^{(0)} - \Phi\left(u^{(0)}\right)\right\| \leq \eta$;
2. *there exists an open domain* $\Omega \subseteq \mathbb{R}^{N-1}$ *such that for each* $u', u'', u''' \in \Omega$ *the following three estimations hold*

   a. $\left\|[F(u', u'')]^{-1}\right\| = \left\|[E - \Phi(u', u'')]^{-1}\right\| \leq B$;
   b. $\left\|\Phi(u', u'')\right\| \leq M$;
   c. $\left\|\Phi(u', u'') - \Phi(u'', u''')\right\| \leq K\left\|u' - u'''\right\|$,

   where B, M, K *are constants;*
3. $h = C_2B^2KM\eta < 1$;
4. *the set* Ω *fully contains a closed ball*

$$\left\|u - u^{(0)}\right\| \leq R, \tag{12}$$

where $R = \dfrac{C_1S_0}{C_2BKM}$, $S_k = \sum\limits_{n=k}^{\infty} h^{2^n}$, $C_1 = 1 + B^2KM\hat{F}$, $C_2 = 1 + B^2K\hat{F}$ *and* $\hat{F} = \sup\limits_{x\in\Omega} \|F(x)\|$.

Then (a) all elements of the sequence $\left(u^{(k)}\right)$*, defined by the method (10)–(11) which starts from the certain* $u^{(0)}$*, lie in the ball (12), (b) the sequence* $\left(u^{(k)}\right)$ *has a limit* u^* *in the ball (12), and (c) the estimation takes place*

$$\left\|u^* - u^{(k)}\right\| \leq \frac{C_1}{C_2BKM}S_k \quad (k = 0,\ 1,\ 2,\ \ldots).$$

Proof. Directly from formulas (10)–(11) we get

$$\tilde{x}^{(k+1)} - \tilde{x}^{(k)} = -\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\left(F\left(\tilde{x}^{(k)}\right) + F\left(x^{(k)}\right)\right); \tag{13}$$

$$\begin{aligned}
&\tilde{x}^{(k+1)} - \Phi\left(\tilde{x}^{(k)}\right) = F\left(\tilde{x}^{(k)}\right) - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\left(F\left(\tilde{x}^{(k)}\right) + F\left(x^{(k)}\right)\right) \\
&= \left[E - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right] F\left(\tilde{x}^{(k)}\right) - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(x^{(k)}\right) \\
&= -\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\left(E - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right) F\left(\tilde{x}^{(k)}\right) \\
&- \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(x^{(k)}\right) \\
&= -\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\left[\Phi\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) F\left(\tilde{x}^{(k)}\right) + F\left(x^{(k)}\right)\right];
\end{aligned} \tag{14}$$

$$\begin{aligned}
&x^{(k)} - \Phi\left(\tilde{x}^{(k)}\right) = \tilde{x}^{(k)} - \Phi\left(\tilde{x}^{(k)}\right) - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^{(k)}\right) \\
&= F\left(\tilde{x}^{(k)}\right) - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^{(k)}\right) \\
&= \left(E - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right) F\left(\tilde{x}^{(k)}\right) \\
&= -\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\left(E - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right) F\left(\tilde{x}^{(k)}\right) \\
&= -\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} \Phi\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) F\left(\tilde{x}^{(k)}\right).
\end{aligned} \tag{15}$$

According to the Definition (1) and formulas (10)–(11) we get

$$\begin{aligned}
&F\left(x^{(k)}\right) = F\left(\tilde{x}^{(k)}\right) + F\left(x^{(k)}, \tilde{x}^{(k)}\right)\left(x^{(k)} - \tilde{x}^{(k)}\right) = F\left(\tilde{x}^{(k)}\right) - F\left(x^{(k)}, \tilde{x}^{(k)}\right) \\
&\times \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^{(k)}\right) = \left(E - F\left(x^{(k)}, \tilde{x}^{(k)}\right)\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right) \\
&\times F\left(\tilde{x}^{(k)}\right) = \left(F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right. \\
&\left.- F\left(x^{(k)}, \tilde{x}^{(k)}\right)\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right) F\left(\tilde{x}^{(k)}\right) \\
&= -\left(F\left(x^{(k)}, \tilde{x}^{(k)}\right) - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right)\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^{(k)}\right);
\end{aligned} \tag{16}$$

$$\begin{aligned}
&F\left(\tilde{x}^{(k+1)}\right) = F\left(\tilde{x}^{(k)}\right) + F\left(\tilde{x}^{(k+1)}, \tilde{x}^{(k)}\right)\left(\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right) \\
&= -F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\left(x^{(k)} - \tilde{x}^{(k)}\right) + F\left(\tilde{x}^{(k+1)}, \tilde{x}^{(k)}\right)\left(\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right) \\
&= -F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\left(x^{(k)} - \tilde{x}^{(k+1)}\right) + \left(F\left(\tilde{x}^{(k+1)}, \tilde{x}^{(k)}\right) - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right) \\
&\times \left(\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right) = -F\left(x^{(k)}\right) - \left(F\left(\tilde{x}^{(k+1)}, \tilde{x}^{(k)}\right) - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right) \\
&\times \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\left(F\left(\tilde{x}^{(k)}\right) + F\left(x^{(k)}\right)\right).
\end{aligned} \tag{17}$$

Based on conditions (a)–(c) the following inequalities could be derived from (13)–(17)

$$\begin{aligned}
&\left\|x^{(k)} - \Phi\left(\tilde{x}^{(k)}\right)\right\| \leqslant \left\|\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right\| \\
&\times \left\|\Phi\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right\| \cdot \left\|F\left(\tilde{x}^{(k)}\right)\right\| \leqslant BM \left\|F\left(\tilde{x}^{(k)}\right)\right\|;
\end{aligned}$$

$$\begin{aligned}
&\left\|F\left(x^{(k)}\right)\right\| \leqslant \left\|F\left(x^{(k)}, \tilde{x}^{(k)}\right) - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right\| \left\|\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right\| \\
&\times \left\|F\left(\tilde{x}^{(k)}\right)\right\| \leqslant KB \left\|x^{(k)} - \Phi\left(\tilde{x}^{(k)}\right)\right\| \cdot \left\|F\left(\tilde{x}^{(k)}\right)\right\| \leqslant B^2KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2;
\end{aligned} \tag{18}$$

$$\begin{aligned}
&\left\|\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right\| \leqslant \left\|\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right\| \cdot \left(\left\|F\left(\tilde{x}^{(k)}\right)\right\| + \left\|F\left(x^{(k)}\right)\right\|\right) \\
&\leqslant B \left\|F\left(\tilde{x}^{(k)}\right)\right\| + B^3KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2;
\end{aligned} \tag{19}$$

$$\begin{aligned}
&\left\|\tilde{x}^{(k+1)} - \Phi\left(\tilde{x}^{(k)}\right)\right\| \leqslant \left\|\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right\| \\
&\times \left(\left\|\Phi\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right\| \cdot \left\|F\left(\tilde{x}^{(k)}\right)\right\| + \left\|F\left(x^{(k)}\right)\right\|\right) \\
&\leqslant B\left(M \left\|F\left(\tilde{x}^{(k)}\right)\right\| + \left\|F\left(x^{(k)}\right)\right\|\right) \leqslant BM \left\|F\left(\tilde{x}^{(k)}\right)\right\| + B^3KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2;
\end{aligned} \tag{20}$$

$$\begin{aligned}
&\left\|F\left(\tilde{x}^{(k+1)}\right)\right\| \leqslant \left\|F\left(x^{(k)}\right)\right\| + \left\|F\left(\tilde{x}^{(k+1)}, \tilde{x}^{(k)}\right) - F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right\| \\
&\times \left\|\left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1}\right\| \cdot \left(\left\|F\left(\tilde{x}^{(k)}\right)\right\| + \left\|F\left(x^{(k)}\right)\right\|\right) \\
&\leqslant B^2KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2 + BK \left\|\tilde{x}^{(k+1)} - \Phi\left(\tilde{x}^{(k)}\right)\right\| \cdot \left(\left\|F\left(\tilde{x}^{(k)}\right)\right\| + \left\|F\left(x^{(k)}\right)\right\|\right) \\
&\leqslant B^2KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2 + BK\left(BM \left\|F\left(\tilde{x}^{(k)}\right)\right\| + B^3KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2\right) \\
&\times \left(\left\|F\left(\tilde{x}^{(k)}\right)\right\| + B^2KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2\right) \\
&\leqslant 2B^2KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2 + B^4K^2M(M+1) \left\|F\left(\tilde{x}^{(k)}\right)\right\|^3 + B^6K^3M^2 \left\|F\left(\tilde{x}^{(k)}\right)\right\|^4.
\end{aligned} \tag{21}$$

Let us consider inequalities (19), (20) and (21). There exist such constants C_1, C_2 and C_3 that

$$\left\|\tilde{x}^{(k+1)} - \tilde{x}^{(k)}\right\| \leqslant C_1B \left\|F\left(\tilde{x}^{(k)}\right)\right\|, \tag{22}$$

$$\left\|\tilde{x}^{(k+1)} - \Phi\left(\tilde{x}^{(k)}\right)\right\| \leqslant C_3BM \left\|F\left(\tilde{x}^{(k)}\right)\right\|, \tag{23}$$

$$\left\|F\left(\tilde{x}^{(k+1)}\right)\right\| \leqslant C_2B^2KM \left\|F\left(\tilde{x}^{(k)}\right)\right\|^2. \tag{24}$$

Using induction we derive from (22), (23) and (24) the following

$$\left\|\widetilde{x}^{(k+1)} - \widetilde{x}^{(k)}\right\| \leqslant \frac{C_1h^{2^k}}{C_2BKM}; \quad \left\|\widetilde{x}^{(k+1)} - \Phi\left(\widetilde{x}^{(k)}\right)\right\| \leqslant \frac{C_3h^{2^k}}{C_2BK}, \tag{25}$$

$$\left\| F\left(\widetilde{x}^{(k+1)}\right)\right\| \leqslant h^{2^{k+1}-1}\eta,$$

where $k = 0,\ 1,\ 2,\ \ldots.$

Based on (25) we obtain

$$\left\|\tilde{x}^{(k+p)} - \tilde{x}^{(k)}\right\| \leqslant \frac{C_1}{C_2BKM}\left(h^{2^k} + h^{2^{k+1}} + \cdots + h^{2^{k+p-1}}\right). \tag{26}$$

Passing to the limit ($p \to \infty$) we obtain

$$\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\| \leqslant \frac{C_1}{C_2BKM}\sum_{n=k}^{\infty} h^{2^n} = \frac{C_1}{C_2BKM}S_k \quad (k = 0,\ 1,\ 2,\ \ldots).$$

Passing to the limit ($k \to \infty$) we see that $\tilde{x}^*$ is a solution to the equation $F(u) = 0$.

Let us show that the elements $\tilde{x}^0$, $\Phi\left(\tilde{x}^0\right)$, $\tilde{x}^1$, $\Phi\left(\tilde{x}^1\right)$, $\ldots$, $\tilde{x}^k$, $\Phi\left(\tilde{x}^k\right)$ belong to the ball (12). This follows from inductively provable inequalities

$$\left\|\tilde{x}^{(k)} - \tilde{x}^{(0)}\right\| \leqslant \left\|\tilde{x}^{(k)} - \tilde{x}^{(k-1)}\right\| + \left\|\tilde{x}^{(k-1)} - \tilde{x}^{(k-2)}\right\| + \ldots$$
$$+ \left\|\tilde{x}^{(1)} - \tilde{x}^{(0)}\right\| \leqslant \frac{C_1h^{2^{k-1}}}{C_2BKM} + \cdots + \frac{C_1h^{2^0}}{C_2BKM} \leqslant \frac{C_1}{C_2BKM}\sum_{n=0}^{k-1} h^{2^n}$$

and

$$\left\|\Phi\left(\tilde{x}^{(k)}\right) - \tilde{x}^{(0)}\right\| \leqslant \left\|\Phi\left(\tilde{x}^{(k)}\right) - \tilde{x}^{(k+1)}\right\| + \left\|\tilde{x}^{(k+1)} - \tilde{x}^{(0)}\right\|$$
$$\leqslant \frac{C_2h^{2^k}}{C_2BK} + \frac{C_1}{C_2BKM}\sum_{n=0}^{k} h^{2^n}. \tag{27}$$

From inequality (27) when $k \to \infty$ we obtain the estimate

$$\left\|\tilde{x}^* - \tilde{x}^{(0)}\right\| \leqslant \frac{C_1S_0}{C_2BKM}$$

and make sure that $\tilde{x}^*$ belongs to the ball (12). □

Let us remark that, if parameters μ and λ of method (10)–(11) lay in the segment $(0,\ 1)$ then condition (3) of Theorem 1 should be rewritten in the following form $h = C_2B^2KM\dfrac{\mu}{\lambda}\eta < 1$. The technique of proof is preserved.

The following theorem gives conditions for convergence with the third order.

Theorem 2. *Assume that*

1. *equation* $F(u) = 0$ *has a solution in the ball*

$$\left\|u - u^{(0)}\right\| \leq \rho; \tag{28}$$

2. *for each u', u'', u''' from the ball*

$$\left\| u - u^{(0)} \right\| \le (1+\alpha)\,\rho \tag{29}$$

the following three estimations hold

$$\begin{array}{ll} a. & \left\| [F(u',u'')]^{-1} \right\| = \left\| [E - \Phi(u',u'')]^{-1} \right\| \le B; \\ b. & \left\| \Phi(u',u'') \right\| \le M; \\ c. & \left\| \Phi(u',u'') - \Phi(u'',u''') \right\| \le K \left\| u' - u''' \right\|, \end{array}$$

where B, M, K are constants, at that $\alpha = \max\{l^2\rho^2,\ M\}$, where $l = \sqrt{2C}BKM$, C is a constant;

3. $l\rho < 1$.

Then (a) the solution u^* of equation $F(u) = 0$ is unique in the ball (28), (b) sequence $\left(u^{(k)}\right)$, defined by the method (10)–(11), converges to u^* with the third order, i.e. the following estimation of the convergence rate holds

$$\left\| u^* - u^{(k)} \right\| \le \frac{1}{l}\,(l\rho)^{3^k} \quad (k = 0,\ 1,\ 2,\ \ldots). \tag{30}$$

Proof. Let $k = 0$, then the estimate (30) is valid on the basis of condition (1) of the theorem. Elements $\tilde{x}^{(0)}$ $\Phi\left(\tilde{x}^{(0)}\right)$ belong to the ball (29), because

$$\begin{aligned} &\left\| \Phi\left(\tilde{x}^{(0)}\right) - \tilde{x}^{(0)} \right\| = \left\| \Phi\left(\tilde{x}^{(0)}\right) - \Phi\left(\tilde{x}^{*}\right) + \tilde{x}^* - \tilde{x}^{(0)} \right\| \\ \leqslant\ &\left\| \Phi\left(\tilde{x}^*, \tilde{x}^{(0)}\right) \right\| \cdot \left\| \tilde{x}^* - \tilde{x}^{(0)} \right\| + \left\| \tilde{x}^* - \tilde{x}^{(0)} \right\| \leqslant (1+M)\,\rho \leqslant (1+\alpha)\,\rho. \end{aligned}$$

Let us analyze the behavior of the error using (10)–(11)

$$\begin{aligned} \tilde{x}^* - \tilde{x}^{(k+1)} &= \tilde{x}^* - x^{(k)} + \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(x^{(k)}\right) \\ &= \tilde{x}^* - x^{(k)} - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} \left(F\left(\tilde{x}^*\right) - F\left(x^{(k)}\right)\right) \\ &= \tilde{x}^* - x^{(k)} - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^*, x^{(k)}\right)\left(\tilde{x}^* - x^{(k)}\right) \\ &= \left(E - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^*, x^{(k)}\right)\right)\left(\tilde{x}^* - x^{(k)}\right) \\ &= \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} \left(F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) - F\left(\tilde{x}^*, x^{(k)}\right)\right)\left(\tilde{x}^* - x^{(k)}\right), \end{aligned}$$

$$\begin{aligned} \tilde{x}^* - x^{(k)} &= \tilde{x}^* - \tilde{x}^{(k)} + \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^{(k)}\right) \\ &= \tilde{x}^* - \tilde{x}^{(k)} - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} \left(F\left(\tilde{x}^*\right) - F\left(\tilde{x}^{(k)}\right)\right) \\ &= \tilde{x}^* - \tilde{x}^{(k)} - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^*, \tilde{x}^{(k)}\right)\left(\tilde{x}^* - \tilde{x}^{(k)}\right) \\ &= \left(E - \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} F\left(\tilde{x}^*, \tilde{x}^{(k)}\right)\right)\left(\tilde{x}^* - \tilde{x}^{(k)}\right) \\ &= \left[F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right)\right]^{-1} \left(F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) - F\left(\tilde{x}^*, \tilde{x}^{(k)}\right)\right)\left(\tilde{x}^* - \tilde{x}^{(k)}\right) \end{aligned}$$

and

$$\tilde{x}^* - \Phi\left(\tilde{x}^{(k)}\right) = \Phi\left(\tilde{x}^*\right) - \Phi\left(\tilde{x}^{(k)}\right) = \Phi\left(\tilde{x}^*, \tilde{x}^{(k)}\right)\left(\tilde{x}^* - \tilde{x}^{(k)}\right).$$

From here we derive that

$$\begin{aligned}
&\left\|F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) - F\left(\tilde{x}^*, x^{(k)}\right)\right\| \\
&= \left\|F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) - F\left(\Phi\left(\tilde{x}^{(k)}\right), \tilde{x}^*\right) + F\left(\Phi\left(\tilde{x}^{(k)}\right), \tilde{x}^*\right) - F\left(\tilde{x}^*, x^{(k)}\right)\right\| \\
&\leqslant \left\|F\left(\tilde{x}^{(k)}, \Phi\left(\tilde{x}^{(k)}\right)\right) - F\left(\Phi\left(\tilde{x}^{(k)}\right), \tilde{x}^*\right)\right\| \\
&+ \left\|F\left(\Phi\left(\tilde{x}^{(k)}\right), \tilde{x}^*\right) - F\left(\tilde{x}^*, x^{(k)}\right)\right\| \leqslant K\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\| \\
&+ K\left\|\Phi\left(\tilde{x}^{(k)}\right) - x^{(k)}\right\| \leqslant K\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\| + K\left\|\tilde{x}^* - \Phi\left(\tilde{x}^{(k)}\right)\right\| + K\left\|\tilde{x}^* - x^{(k)}\right\| \\
&\leqslant K(M+1)\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\| + K\left\|\tilde{x}^* - x^{(k)}\right\|,
\end{aligned}$$

$$\left\|\tilde{x}^* - x^{(k)}\right\| \leqslant BK\left\|\tilde{x}^* - \Phi\left(\tilde{x}^{(k)}\right)\right\| \cdot \left\|\tilde{x}^* - \tilde{x}^{(k)}\right\| \leqslant BKM\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\|^2$$

and

$$\begin{aligned}
&\left\|\tilde{x}^* - \tilde{x}^{(k+1)}\right\| \leqslant B^2KM\left(K(M+1)\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\| + BK^2M\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\|^2\right) \\
&\times \left\|\tilde{x}^* - \tilde{x}^{(k)}\right\|^2 \leqslant B^2K^2M(M+1)\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\|^3 + B^3K^3M^2\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\|^4.
\end{aligned}$$

There is a positive constant C such that

$$\left\|\tilde{x}^* - \tilde{x}^{(k+1)}\right\| \leqslant CB^2K^2M(M+1)\left\|\tilde{x}^* - \tilde{x}^{(k)}\right\|^3, \tag{31}$$

that means the method (10)–(11) converges with at least third order. From (31) by induction we obtain the estimate (30).

The estimate (31) is valid under the assumption that the elements $\tilde{x}^{(k)}$ and $\Phi\left(\tilde{x}^{(k)}\right)$ belong to the ball. Let us show that this is indeed so. For $k \geqslant 1$ we have

$$\left\|\tilde{x}^{(k)} - \tilde{x}^{(0)}\right\| \leqslant \left\|\tilde{x}^{(k)} - \tilde{x}^*\right\| + \left\|\tilde{x}^* - \tilde{x}^{(0)}\right\| \leqslant \frac{1}{l}(l\rho)^{3^k} + \rho \leqslant \left(1 + l^2\rho^2\right)\rho \leqslant (1+\alpha)\rho$$

and

$$\begin{aligned}
&\left\|\Phi\left(\tilde{x}^{(k)}\right) - \tilde{x}^{(0)}\right\| \leqslant \left\|\Phi\left(\tilde{x}^{(k)}\right) - \Phi\left(\tilde{x}^*\right) + \tilde{x}^* - \tilde{x}^{(0)}\right\| \\
&\leqslant \left\|\Phi\left(\tilde{x}^*, \tilde{x}^{(k)}\right)\left(\tilde{x}^* - \tilde{x}^{(k)}\right)\right\| + \left\|\tilde{x}^* - \tilde{x}^{(0)}\right\| \\
&\leqslant Ml^2\rho^3 + \rho < (1+M)\rho \leqslant (1+\alpha)\rho.
\end{aligned}$$

From (31) and what was deduced above it follows that $\lim \tilde{x}^{(k)} = \tilde{x}^*$ as $k \to \infty$. If the equation would have a solution in ball (28), then, using similar reasoning, we could show that $\lim \tilde{x}^{(k)} = \tilde{x}^{**}$ as $k \to \infty$. Based on the uniqueness of the limit element of a convergent sequence $\left(\tilde{x}^{(k)}\right)$ the solution $\tilde{x}^*$ is unique in ball (28). □

The Theorems 1 and 2 impose less stringent conditions on the operator of the first divided difference than is done, for example, in [1].

4 Numerical Examples

Let us consider a concrete example. Namely, in Eq. (1) we take $\omega(u) = exp(u)$ and perform numerical simulations in various time and space grids.

Let us consider the initial boundary value problem

$$\frac{\partial e^u}{\partial t} = \frac{\partial^{1.5} u}{\partial |x|^{1.5}} - f(x,t) \tag{32}$$

on the domain $x \in (0, 1),\ t \in (0, 4\pi)$, where

$$\begin{aligned} f(x,t) = &- \exp\left(x^2\,(1-x)^2\,\cos(t)\right) x^2\,(1-x)^2\,\sin(t) \\ &- \frac{\cos(t)}{\sqrt{2\pi}}\,\frac{16}{315}\,\Big(32\,(1-x)^{2.5} + 32\,x^{2.5} + 160\,x^{1.5}\,(-1.125+x) \\ &- 5\,(1-x)^{1.5}\,(4+32\,x) + 60\,x^{0.5}\,(1.3125 - 2.25\,x + x^2) \\ &+ 3.75\,(1-x)^{0.5}\,(1+4\,x+16\,x^2)\Big). \end{aligned}$$

Initial and boundary conditions are defined as follow

$$u(x,0) = x^2\,(1-x)^2, \quad 0 \le x \le 1,$$

$$u(0,t) = 0,\ u(1,t) = 0, \quad 0 \le t \le 4\pi.$$

Problem (32) has an exact solution $u(x,t) = x^2\,(1-x)^2\,\cos t$.

The algorithm was implemented using Python 3.7, all computations were performed in a double precision. To compare the Newton method and the elaborated method (10)–(11) we took the accuracy which should be achieved during the iterations to be $\epsilon = 10^{-5}$, i.e. $\|u_{j+1}^{(k)} - u_{j+1}^{(k-1)}\| \le \epsilon$.

Results of numerical experiments with method (10)–(11) are presented in Table 1. The third column shows the maximum of absolute difference between the exact and numerical solutions $\mathbf{diff}_{\Delta,h} = \max\limits_{i,j} |u_j^i - u(x_i, t_j)|$, $i = 0, \ldots, N$, $j = 0, \ldots, M$, where N and M are the number of segments in space and time. The fourth column represents the ratio of the error reduction as the space grid refined.

Table 1 Convergence of method (10)–(11) in various time and space grids

$\varDelta$	h	$\mathrm{diff}_{\varDelta,h}$	Error rate
$\pi/10$	1×2^{-2}	9.5399×10^{-3}	–
	1×2^{-3}	2.7679×10^{-3}	3.4465
	1×2^{-4}	1.2202×10^{-3}	2.2684
	1×2^{-5}	8.6654×10^{-4}	1.4081
	1×2^{-6}	7.8471×10^{-4}	1.1043
$\pi/20$	1×2^{-2}	8.9632×10^{-3}	–
	1×2^{-3}	2.2025×10^{-3}	4.0694
	1×2^{-4}	6.5309×10^{-4}	3.3725
	1×2^{-5}	2.9702×10^{-4}	2.1988
	1×2^{-6}	2.1477×10^{-4}	1.3829
$\pi/40$	1×2^{-2}	8.8006×10^{-3}	–
	1×2^{-3}	2.0554×10^{-3}	4.2816
	1×2^{-4}	5.0953×10^{-4}	4.0339
	1×2^{-5}	1.5430×10^{-4}	3.3021
	1×2^{-6}	7.2263×10^{-5}	2.1353

In the series of experiments with $\varDelta = \pi/40$ the error related to the time discretization is small in comparison with the error related to the coordinate discretization; the analysis of the error behavior reveals the second convergence with respect to space variables, i.e. when the step becomes half as much, the error becomes almost two times less as well.

The analysis of the data in the table shows that only the consistent decrease of steps yields the decrease of error. Indeed, in the series of experiments with $\varDelta = \pi/10$ the halving of h does not cause the corresponding decrease of error, because the total error is mostly induced by the time discretization.

To show the advantages of the developed method over the Newton method we compare the number of iterations necessary to achieve the desired accuracy, see Table 2. The row "Iteration number" represents the average number of iterations done at each time layer.

Definition 2 **([13]).** The efficiency index of the iterative method is called quantity $p^{1/\theta}$, where p is the order of convergence of the method, θ is the number of function calculations at each iterative step.

The efficiency index of method (10)–(11) is $3^{1/2} \approx 1.732$ while Newton method has the efficiency index 2 only. So, despite the fact that the elaborated method (10)–(11) requires two steps per iterations and, therefore, two calculations of function values, the total amount of computational work is fewer than in Newton method.

Table 2 Average number of iterations and errors norms

N	16	16	16	16
M	10	20	40	80
	Method (10)–(11)			
diff	1.2202×10^{-3}	6.5309×10^{-4}	5.0953×10^{-4}	4.978×10^{-4}
Iter. num.	1.63	1.50	1.06	1.03
	Newton method			
diff	1.1971×10^{-3}	6.734×10^{-4}	6.1102×10^{-4}	5.012×10^{-4}
Iter. num.	5.125	5.125	5.03	4.53

Acknowledgements This work was supported by RFBR Grant 19-01-00019. The work of Iumanova was also supported by Act 211 Government of the Russian Federation, contract No 02.A03.21.0006.

References

1. Amat, S., Busquier, S., Bermudez, C., Magrenan, A.: Expanding the applicability of a third order Newton-type method free of bilinear operators. Algorithms **8**, 669–679 (2015). https://doi.org/10.3390/a8030669
2. Berinde, V.: Iterative Approximation of Fixed Points. Springer (2007)
3. Brezinski, C.: Convergence acceleration during the 20th century. J. Comput. Appl. Math. **122**, 1–21 (2000)
4. De Staelen, R., Hendy, A.: Numerically pricing double barrier options in a time-fractional Black-Scholes model. Comput. Math. Appl. (2019). https://doi.org/10.1016/j.camwa.2017.06.005
5. Flores, S., Macias-Diaz, J.E., Hendy, A.: Discrete monotone method for space-fractional nonlinear reaction-diffusion equations. Adv. Diff. Eq. (2019). https://doi.org/10.1186/s13662-019-2267-1
6. Gorbova, T.V., Pimenov, V.G., Solodushkin, S.I.: Difference schemes for the nonlinear equations in partial derivatives with heredity. In: Dimov, I., Farago, I., Vulkov, L. (eds) Finite Difference Methods. Theory and Applications, FDM 2018. Lecture Notes in Computer Science, vol. 11386. pp. 258–265. Springer, Cham (2019)
7. Kilbas, A.A., Srivastava, H.M., Trujillo, J.J.: Theory and Applications of Fractional Differential Equations. Elsevier, Amsterdam (2006)
8. Pimenov, V., Hendy, A.: An implicit numerical method for the solution of the fractional advection-diffusion equation with delay. Trudy Instituta Matematiki i Mekhaniki UrO RAN (2016). https://doi.org/10.1007/s001090000086
9. Pimenov, V., Hendy, A.: A fractional analog of Crank-Nicholson method for the two sided space fractional partial equation with functional delay. Ural Math. J. (2016). https://doi.org/10.15826/umj.2016.1.005
10. Pimenov, V.G.: General linear methods for the numerical solution of functional-differential equations. Diff. Eq. **37**(1), 116–127 (2001)
11. Samarskii, A.A.: The Theory of Difference Schemes. Taylor & Francis Inc., New York (2001)
12. Srivastava, V.K., Kumar, S., et al.: Two-dimensional time fractional-order biological population model and its analytical solution. Egypt J. Basic Appl. Sci. **1**, 71–76 (2014)
13. Traub, J.F.: Iterative Methods for the Solution of Equations. AMS (1982)

14. Ul'm, S.Yu.: On generalized divided differenes. I, Izv. Akad. Nauk Est. SSR, Fiz.-Mat. **16**, 13–26 (1967)
15. Wang, X., Liu, F., Chen, X.: Novel second-order accurate implicit numerical methods for the Riesz space distributed-order advection-dispersion equations. Adv. Math. Phys. **4**, 1–14 (2015)
16. Zhou, Y., Luo, Z.: A Crank–Nicolson finite difference scheme for the Riesz space fractional-order parabolic-type sine-Gordon equation. Adv. Differ. Eq. **2018** (2018)

Non-homogeneous Boundary Problems for One-Dimensional Flow of the Compressible Viscous and Heat-Conducting Micropolar Fluid

Ivan Dražić

Abstract We consider nonstationary 1-D flow of a compressible viscous and heat-conducting micropolar fluid which is in the thermodynamical sense perfect and polytropic.

In the first part of the work we present corresponding initial-boundary value problems whereby we allow non-homogeneous boundary conditions for velocity, microrotation or temperature.

In the second part of the work we present existence results for described problems under the additional assumption that the initial density and initial temperature are strictly positive.

1 Introduction

Today, modern engineering is shifting its focus from macro to micro level. Nanotechnology is slowly entering all spheres of human activity and it is up to mathematics to properly follow this trend with associated models. Continuum mechanics have been macro-oriented for a long time, and any attempt to model a phenomenon on a microscale has generally resulted in a model that was so complex that its mathematical analysis would generally not be possible. That was until the 1960s when Eringen established the micropolar continuum model with the introduction of a new variable, microrotation. With this new variable he was able to model phenomena at the micro level.

In this paper, we concentrate on an isotropic compressible and thermally conductive micropolar fluid, which we assume to be perfect and polytropic in thermodynamic terms. The mathematical analysis of this model began with a one-dimensional case in [4], and that case is the focus of this paper. In addition to the one-dimensional case, a multidimensional case with different symmetries in solutions has been considered to date, and we refer to [1] for details.

I. Dražić (⌧)
Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia
e-mail: idrazic@riteh.hr

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_30

In the last few years, the micropolar fluid model is successfully applied in different engineering areas, as well as in the field of medicine. For some specific applications, we refer to [2], together with references cited herein.

The main goal of this paper is to give an overview of recent results concerning the described one-dimensional model in the relation to different boundary conditions, with the focus on the existence of the solution of the associated system of partial differential equations.

2 The Mathematical Model

In [4] the following system was established:

$$\rho_t + \rho^2 v_x = 0, \tag{1}$$

$$v_t = (\rho v_x)_x - K(\rho\theta)_x, \tag{2}$$

$$\rho\omega_t = A\,(\rho(\rho\omega_x)_x - \omega), \tag{3}$$

$$\rho\theta_t = -K\rho^2\theta v_x + \rho^2(v_x)^2 + \rho^2(\omega_x)^2 + \omega^2 + D\rho(\rho\theta_x)_x, \tag{4}$$

which describes the one-dimensional motion of an isotropic, viscous and heat conducting micropolar fluid, which is in the thermodynamical sense perfect and polytropic. Equations (1)-(4) are, respectively, local forms of conservation laws for the mass, momentum, momentum moment and energy, where we have the following notations:

- ρ - mass density,
- v - velocity,
- ω - microrotation velocity,
- θ - absolute temperature,
- K, A, D - positive constants.

The system (1)-(4) is given in the Lagrangian form and considered in the domain $]0,1[\times\mathbf{R}^+$, together with smooth enough initial conditions:

$$\rho(x,0) = \rho_0(x), \quad v(x,0) = v_0(x), \quad \omega(x,0) = \omega_0(x), \quad \theta(x,0) = \theta_0(x), \tag{5}$$

whereby we assume that the functions ρ_0 and θ_0 are strictly positive and bounded:

$$m \le \rho_0(x), \theta_0(x) \le M, \quad x \in]0,1[, \tag{6}$$

where $m, M \in \mathbf{R}^+$.

To complete the problem, we will propose various kinds of boundary conditions and consider the properties of the so-called generalized solution which is given in the following definition.

Definition 1. Given any $T \in \mathbf{R}^+$, a generalized solution to the system (1)-(4) in the domain $Q_T =]0,1[\times]0,T[$, together with appropriate initial and boundary conditions is a function

$$(x,t) \mapsto (\rho, v, \omega, \theta)(x,t), \quad (x,t) \in Q_T, \tag{7}$$

where

$$\rho \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(]0,1[)) \cap \mathrm{H}^1(Q_T)\ ,\ \inf_{Q_T} \rho > 0\ , \tag{8}$$

$$v,\omega,\theta \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(]0,1[)) \cap \mathrm{H}^1(Q_T) \cap \mathrm{L}^2(0,T;\mathrm{H}^2(]0,1[)), \tag{9}$$

that satisfies the Eqs. (1)-(4) a.e. in Q_T and initial and boundary conditions in the sense of traces.

Let us mention that by using the embedding and interpolation theorems one can conclude that our generalized solution could be treated as a strong solution. In fact, we have

$$\rho \in \mathrm{L}^\infty(0,T;\mathrm{C}([0,1])) \cap \mathrm{C}([0,T],\mathrm{L}^2(]0,1[))\ , \tag{10}$$

$$v,\omega,\theta \in \mathrm{L}^2(0,T;\mathrm{C}^1([0,1])) \cap \mathrm{C}([0,T],\mathrm{H}^1(]0,1[)), \tag{11}$$

$$v,\omega,\theta \in \mathrm{C}(\overline{Q}_T). \tag{12}$$

3 Existence of the Solution in Dependence of Boundary Conditions

3.1 *Homogeneous Case*

Let us first consider the problem with homogeneous boundary conditions for velocity, microrotation and heat-flux:

$$v(0,t) = v(1,t) = 0, \quad \omega(0,t) = \omega(1,t) = 0, \quad \theta_x(0,t) = \theta_x(1,t) = 0. \tag{13}$$

This conditions were analyzed in [4] and [5], from where we have the following theorem:

Theorem 1. *Let the functions* $\rho_0, \theta_0 \in \mathrm{H}^1(]0,1[)$ *satisfy the conditions (6) and let* $v_0, \omega_0 \in \mathrm{H}_0^1(]0,1[)$. *Then for any* $T \in \mathbf{R}^+$ *there exists unique generalized solution to the problem (1)-(5) and (13) in the domain* Q_T *having the property*

$$\theta > 0 \ \ in\ \overline{Q}_T. \tag{14}$$

Using the Faedo-Gelerikin method, local existence was proved in [4] as well as the uniqueness of the solution. Then, based on extension principle and series of a priori estimates, in [5] the global existence theorem was established.

From physical point of view, these boundary conditions are used to model the flow between solid thermo-insulated walls.

For this case we also know that the solution is regular and exponentially stable; for recent progress in this case we refer to [3].

3.2 Non-homogeneous Conditions for Velocity and Microrotation

Now we will consider the problem with homogeneous boundary conditions for heat-flux, but with inhomogeneous boundary conditions for velocity and microrotation:

$$v(0,t) = \mu_0(t),\ v(1,t) = \mu_1(t),\ \omega(0,t) = \nu_0(t),\ \omega(1,t) = \nu_1(t), \tag{15}$$

$$\theta_x(0,t) = \theta_x(1,t) = 0. \tag{16}$$

For these condition we have the following theorem:

Theorem 2. *Let the functions* $\rho_0, \theta_0, v_0, \omega_0 \in \mathrm{H}^1(]0,1[)$ *and* $\mu_0, \mu_1, \nu_0, \nu_1 \in \mathrm{H}^2(]0,T[)$ *satisfy the conditions (6) as well as the compatibility conditions:*

$$v_0(0) = \mu_0(0), \tag{17}$$

$$v_0(1) = \mu_1(0), \tag{18}$$

$$\omega_0(0) = \nu_0(0), \tag{19}$$

$$\omega_0(1) = \nu_1(0). \tag{20}$$

Let also exist a constant $\delta > 0$ *such that*

$$l(t) = \int_0^1 \frac{1}{\rho(0)}dx + \int_0^t [\mu_1(\tau) - \mu_0(\tau)]\,d\tau \geq \delta, \tag{21}$$

for $t \in]0, T[$. Then for any $T \in \mathbf{R}^+$ there exists a generalized solution to the problem (1)-(5) and (15)-(16) in the domain Q_T having the property

$$\theta > 0 \;\; in\; \overline{Q}_T. \tag{22}$$

The local existence for this case was proved in [8], while global existence is given in [10]. The proofs in these two works are very similar as the ones in [4] and [5], but here the procedure of homogenization of boundary conditions is performed first which, produces much complex system comparing the case with homogeneous boundary conditions. From physical point of view these boundary conditions are used to model the piston problem. Regarding the further mathematical properties, for this case we just know that the solution is regular. For details we refer to [9].

3.3 Non-homogeneous Conditions for Temperature

In this subsection we will consider the problem with homogeneous boundary conditions for velocity and microrotation, but inhomogeneous boundary conditions for temperature:

$$v(0,t) = v(1,t) = 0, \quad \omega(0,t) = \omega(1,t) = 0 \tag{23}$$

$$\theta(0,t) = \mu_0(t), \quad \theta(1,t) = \mu_1(t). \tag{24}$$

For these condition we have the following theorem:

Theorem 3. *Let the functions $\rho_0, \theta_0 \in \mathrm{H}^1(]0,1[)$ satisfy the conditions (6) and let $v_0, \omega_0 \in \mathrm{H}_0^1(]0,1[)$. Let the functions $\mu_0, \mu_1 \in \mathrm{H}^2(]0,T[)$ satisfy the compatibility conditions:*

$$\theta_0(0) = \mu_0(0), \tag{25}$$

$$\theta_0(1) = \mu_1(0), \tag{26}$$

as well as the conditions:

$$\mu_0(t) \geq m, \qquad \mu_1(t) \geq m, \tag{27}$$

for $t \in]0, T[$ and m defined in (6). Then for any $T \in \mathbf{R}^+$ there exists a generalized solution to the problem (1)-(5) and (23)-(24) in the domain Q_T having the property

$$\theta > 0 \;\; in\; \overline{Q}_T. \tag{28}$$

This case was analyzed in [11] and [12], whereby the local existence was established in [11] and global in [12].

From physical point of view, these boundary conditions are used to model the flow between solid walls with varying temperature. Mathematical analysis of the properties of the solution, especially regarding the regularity and stabilization for this case is still open problem.

3.4 The Cauchy Problem

Here we consider the problem (1)-(5) defined for $x \in \mathbf{R}$ without boundary conditions. In this case we have the following theorem.

Theorem 4. *Let the initial functions ρ_0, θ_0, ω_0 and θ_0 satisfy the following conditions:*

$$m \le \rho_0(x), \theta_0(x) \le M, \quad x \in \mathbf{R}, \tag{29}$$

where $m, M \in \mathbf{R}^+$, and

$$\rho_0 - 1, \theta_0 - 1, v_0, \omega_0 \in \mathrm{H}^1(\mathbf{R}). \tag{30}$$

Then for any $T \in \mathbf{R}^+$ there exists an unique solution to the problem (1)-(5) in the domain $Q_T = \mathbf{R}\times]0, T[$ having the properties:

$$\rho - 1 \in \mathrm{L}^\infty(0, T; \mathrm{H}^1(\mathbf{R})) \cap \mathrm{H}^1(Q_T), \tag{31}$$

$$v, \omega, \theta - 1 \in \mathrm{L}^\infty(0, T; \mathrm{H}^1(\mathbf{R})) \cap \mathrm{H}^1(Q_T) \cap \mathrm{L}^2(0, T; \mathrm{H}^2(\mathbf{R})), \tag{32}$$

that satisfies the equations (1)-(4) a.e. in Q_T and conditions (5) in the sense of traces.

This theorem was proved in [6] and [7]. In [6] the existence was proved, while uniqueness was established in [7]. The uniqueness was proved using the classical approach with Gronwall inequality, and for existence the method of domain extending was used.

For recent results in mathematical analysis of these case we refer to [13].

4 Conclusion

In this paper, we considered the one dimensional flow of a compressible micropolar fluid, whereby we assumed that the fluid is heat-conducting, isotropic and in thermodynamical sense perfect and polytropic. The corresponding system is coupled by smooth enough initial conditions and we analysed different cases of boundary conditions in relation to existence of generalized solution.

Acknowledgements This work has been fully supported by the University of Rijeka, Croatia under the project 17.10.2.2.01 (Initial-boundary problems in the research of the micropolar continuum model) and the project uniri-prirod-18-118-1257 (Analysis of mathematical models of fluid mechanics and technical systems using data-driven algorithms for Koopman operator), as well as trough funding of University of Rijeka Foundation.

References

1. Dražić, I.: 3-D flow of a compressible viscous micropolar fluid model with spherical symmetry: a brief survey and recent progress. Rev. Math. Phys. **30**, 1830001 (2018)
2. Dražić, I.: Dimensionless formulation for the one-dimensional compressible flow of the viscous and heat-conducting micropolar fluid. Phys. Astron. Int. J. **2**(5), 420–423 (2018)
3. Dražić, I., Simčić, L.: One-dimensional flow of a compressible viscous and heat-conducting micropolar fluid with homogeneous boundary conditions: a brief survey of the theory and recent progress. Global Stochast. Anal. **5**(1), 45–55 (2018)
4. Mujaković, N.: One-dimensional flow of a compressible viscous micropolar fluid: a local existence theorem. Glasnik Matematički. **33**, 71–91 (1998)
5. Mujaković, N.: One-dimensional flow of a compressible viscous micropolar fluid: a global existence theorem. Glasnik Matematički. **33**, 199–208 (1998)
6. Mujaković, N.: One-dimensional flow of a compressible viscous micropolar fluid: the Cauchy problem. Math. Commun. **10**(1), 1–14 (2005)
7. Mujaković, N.: Uniqueness of a solution of the Cauchy problem for one-dimensional compressible viscous micropolar fluid model. Appl. Math. E-Notes **6**, 113–118 (2006)
8. Mujaković, N.: Non-homogeneous boundary value problem for one-dimensional compressible viscous micropolar fluid model: a local existence theorem. Ann. Univ. Ferrara Sez. VII Sci. Mat. **53**(2), 361–379 (2007)
9. Mujaković, N.: Nonhomogeneous boundary value problem for one-dimensional compressible viscous micropolar fluid model: regularity of the solution. Bound. Value Probl. **189748**, 1–15 (2008)
10. Mujaković, N.: Non-homogeneous boundary value problem for one-dimensional compressible viscous micropolar fluid model: a global existence theorem. Math. Inequal. Appl. **12**(3), 651–662 (2009)
11. Mujaković, N.: 1-D compressible viscous micropolar fluid model with non-homogeneous boundary conditions for temperature: a local existence theorem. Nonlinear Anal. Real World Appl. **13**(4), 1844–1853 (2012)
12. Mujaković, N.: The existence of a global solution for one dimensional compressible viscous micropolar fluid with non-homogeneous boundary conditions for temperature. Nonlinear Anal. Real World Appl. **19**, 19–30 (2014)
13. Qin, Y., Wang, T., Liu, X.: Global Existence and Uniqueness of Nonlinear Evolutionary Fluid Equations. Birkhäuser, Basel (2015)

The General Case of Cutting GML Bodies: The Geometrical Solution

Johan Gielis, Diego Caratelli, and Ilia Tavkhelidze

Abstract The original motivation to study this class of geometrical objects of Generalized Möbius-Listing *GML* surfaces and bodies was the observation that the solution of boundary value problems greatly depends on the structure of the boundary of domains. Since around 2010 *GML*'s were merged with (continuous) Gielis Transformations, which provide a unifying description of geometrical shapes, as a generalization of the Pythagorean Theorem. The resulting geometrical objects can be used for modeling a wide range of natural shapes and phenomena.

The cutting of *GML* bodies and surfaces, with the Möbius strip as one special case, is related to the field of knots and links, and classifications were obtained for *GML* with cross sectional symmetry of 2, 3, 4, 5 and 6. The general case of cutting *GML* bodies and surfaces, in particular the number of ways of cutting, could be solved by reducing the 3D problem to planar geometry [1]. This also unveiled a range of connections with topology, combinatorics, elasticity theory and theoretical physics.

1 Generalized Möbius-Listing surfaces and bodies

GML_m^n are torus-like surfaces or bodies, which are constructed by identifying opposite sides of a cylinder or prisms with given cross sections (Fig. 1) [1, 2], whereby the original cylinder or prism may be twisted. These cross sections of the GML_m^n bodies are closed planar curves with symmetry *m*. For the classical cylinder or Möbius band, the cross section is a line, swept along a path forming a ribbon and twisted an even or odd number of times around the basic line, which is the line traced out by the centre of the line. The lower index *m* in GML_m^n determines the symmetry of the

J. Gielis (✉)
Department of Biosciences Engineering, University of Antwerp, Antwerp, Belgium
e-mail: johan.gielis@uantwerpen.be

D. Caratelli
The Antenna Company and Eindhoven University of Technology, Eindhoven, The Netherlands

I. Tavkhelidze
Department of Mathematics, Tbilisi State University, Tbilisi, Georgia

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_31

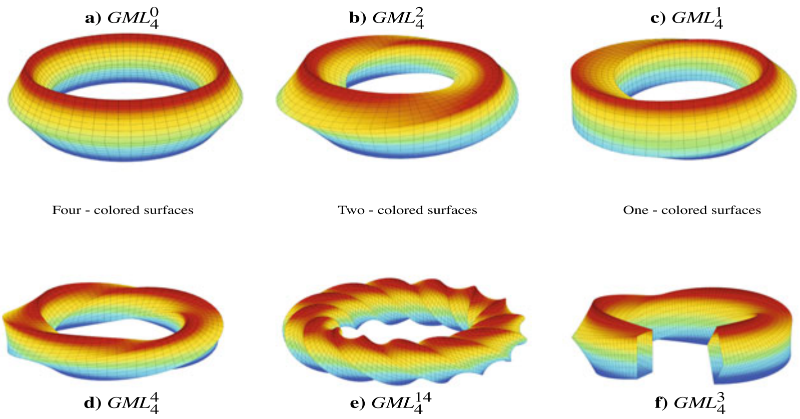


Fig. 1 GML surfaces and bodies

cross section and the upper index *n* describes the number of twists. *GML* surfaces or bodies can either be closed (e.g. a classical torus) or not (Fig. 1f). We observe in GML_4^n:

1. For GML_4^0 and GML_4^4 (Fig. 1 a&d) there are no ($n = 0$) or four ($n = 4$) twists of the original prism of 90° each. The original identification of opposite sides of the square prism with vertices *ABCD* and *A'B'C'D'* is then *AA'-BB'-CC'-DD'*. This corresponds to four-coloured surfaces of the *GML*, and this is the case for any multiple of four.
2. For GML_4^2 and GML_4^{14} (Fig. 1 b& e) there are two ($n = 2$) or fourteen ($n = 14$) twists of the original prism of 90° each. The original identification of opposite sides of the square prism with vertices *ABCD* and *A'B'C'D'* is then *AC'-BD'-CA'-DB'* and this corresponds to two coloured surfaces. With one brush and two colours, the whole surface can be painted, and this is the case for any multiple of four plus 2.
3. For GML_4^1 and GML_4^3 (Fig. 1 c&f), there are 1 ($n = 1$) and 3 ($n = 3$) twists of the original prism of 90° each. The original identification of opposite sides of the square prism with vertices *ABCD* and A*'B'C'D'* is then *AD'-BA'-CB'-DC'* for $n = 1$. The case $n = 3$ is symmetrical, i.e. one twist of 90° anticlockwise instead of clockwise for $n = 1$ and both lead to one coloured surfaces. With one brush and one colour only, the whole surface can be painted, and this is the case for any multiple of four plus or minus one.

The planar curves with symmetry *m* can be regular polygons or any closed plane curve - including circles. *GML* surfaces are generated when only the curve itself - as boundary of a region is considered (Fig. 1f for GML_4^3), or they are bodies when also the disk enclosed by the curve is considered. A convenient method for both sections and curve is the use of Gielis curves and transformations, a generalization of Lamé curves [3]. These curves have been defined for 3D and higher as well [4], in the spirit of spherical coordinates, but this immediate use in *GML* avoids the laborious generalization to cylindrical or other coordinate systems, since such systems can all be considered as subsets of *GML*. They are also a generalization of canal surfaces defining boundaries and disks.

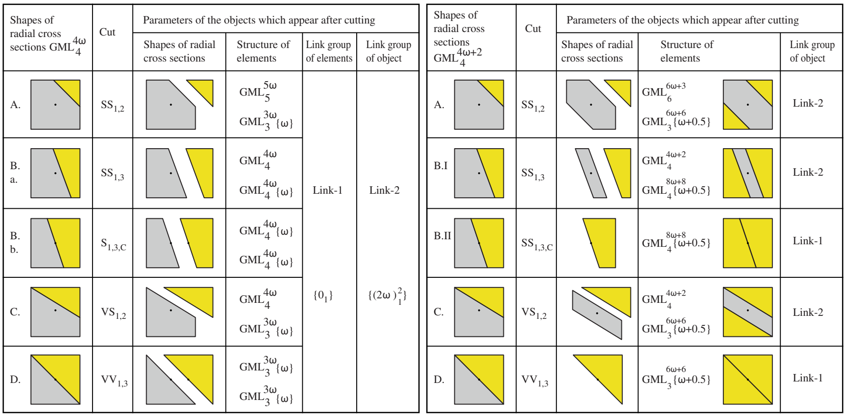
Interestingly, Gielis transformations can morph manifolds in a space to the space itself (and back). This is in the spirit of Gabriel Lamé (1795–1860), which has been characterised as follows [5]: *"The importance to consider all systems of curvilinear coordinates, and not only the four or five (Cartesian, oblique, cylindrical, spherical, bipolar) in use around 1830 is all in all considerable. One can compare this move to Descartes, who instead of about ten curves studied by the Ancients, passes immediately to an infinite number of curves, that can represent physical phenomena. Now, following Lamé, before representing the phenomenon itself, to each physical situations a curvilinear coordinate system is associated, which reflects the shape of the place where it resides"*.

The current investigation is to determine the total number of possible cuts if the *GML* surfaces and bodies are cut along a certain line or surface. This is inspired by the cutting of the original Möbius band and so far, the classification of cutting of GML_m^n surfaces and bodies was achieved for classic Möbius bands [3], and for GML_m^n with m = 2, 3, 4, 5 and 6 [6–9]. A full classification was also achieved for cutting of classic Möbius bands with any k number of (parallel) knives [10]. These classifications revealed a close link between the cutting of *GML* bodies and surfaces, the study of knots and links, and with the colouring of surfaces. The challenge remains to classify the cutting of general *GML* bodies when the cross section of the *GML* body is a regular *m*-polygon, for any value of *m*. However, the choice for regular polygons does not restrict immediate generalizations.

2 The Cutting of GML Bodies

Cutting is performed with 1) a straight knife, which 2) cuts perpendicular to the polygonal cross section of the *GML*, and 3) the knife cuts the *m*-polygon boundary exactly in two points or two times (depending on the thickness of the knife). For 3) there are three possibilities: the cut of the polygon can be from a vertex to a vertex *VV*, from a vertex to a side or edge *VS*, or from side to side *SS* (= edge to edge). The precise orientation of this knife (and the positions where it cuts the boundary) is maintained during the complete cutting process, until the knife returns to its starting position, and the cutting is completed. Depending on the number of twists, a number of independent bodies results, that is related to the divisors of *m*.

If we consider the *GML* with square cross section of Fig. 1, the results of cutting are shown in Fig. 2 (a. $GML_4^{4\omega}$ and b. $GML_4^{4\omega+2}$). Figure 2a shows the results of cutting for $GML_4^{4\omega}$, with ω a natural number (multiples of 4). In the case of $\omega = 0$ the structure is untwisted as GML_4^0 in Fig. 1, but in the case of $\omega = 1$ the structure is twisted 360° as GML_4^4 in Fig. 1. If a cut is made, in all cases, two bodies result. Depending on the cut made, the resulting bodies have a certain cross sectional shape



Shapes of radial cross sections $GML_4^{4\omega}$	Cut	Parameters of the objects which appear after cutting			
		Shapes of radial cross sections	Structure of elements	Link group of elements	Link group of object
A.	$SS_{1,2}$		$GML_5^{5\omega}$ $GML_3^{3\omega}\{\omega\}$	Link-1 $\{0_1\}$	Link-2 $\{(2\omega)_1^2\}$
B. a.	$SS_{1,3}$		$GML_4^{4\omega}$ $GML_4^{4\omega}\{\omega\}$		
B. b.	$S_{1,3,C}$		$GML_4^{4\omega}\{\omega\}$ $GML_4^{4\omega}\{\omega\}$		
C.	$VS_{1,2}$		$GML_4^{4\omega}$ $GML_3^{3\omega}\{\omega\}$		
D.	$VV_{1,3}$		$GML_3^{3\omega}\{\omega\}$ $GML_3^{3\omega}\{\omega\}$		

Shapes of radial cross sections $GML_4^{4\omega+2}$	Cut	Parameters of the objects which appear after cutting		
		Shapes of radial cross sections	Structure of elements	Link group of object
A.	$SS_{1,2}$		$GML_6^{6\omega+3}$ $GML_3^{6\omega+6}\{\omega+0.5\}$	Link-2
B.I	$SS_{1,3}$		$GML_4^{4\omega+2}$ $GML_4^{8\omega+8}\{\omega+0.5\}$	Link-2
B.II	$SS_{1,3,C}$		$GML_4^{8\omega+8}\{\omega+0.5\}$	Link-1
C.	$VS_{1,2}$		$GML_4^{4\omega+2}$ $GML_3^{6\omega+6}\{\omega+0.5\}$	Link-2
D.	$VV_{1,3}$		$GML_3^{6\omega+6}\{\omega+0.5\}$	Link-1

Fig. 2 **a** $GML_4^{4\omega}$ and **b** $GML_4^{4\omega+2}$ [8]

and a defined number of twists, as indicated in the column "*structure of the elements*". For example a cut from side 1 to side 3, results in case B, and if this cut contains the center (SB) case Bb results. In both cases the resulting bodies have quadrangular shapes, and are twisted 4ω times. For a diagonal cut from vertex to vertex (case D) two triangular bodies results, each twisted 3ω times.

The resulting bodies for $GML_4^{4\omega+2}$ bodies (Fig. 1b and d) are shown in Fig. 2b. For case A also two bodies result, a hexagonal one and a triangular one, with the specified number of twists, $GML_6^{6\omega+3}$ and $GML_3^{6\omega+3}$ respectively. For $GML_4^{4\omega+2}$ (i.e. twists of 180°) it is possible that only one body results (cases BII and D), but with a higher number of twists ($GML_4^{8\omega+8}$ for BII and $GML_3^{6\omega+6}$ for D). This happens when the cut is made through the centre and this is completely analogous to the Möbius phenomenon resulting from cutting a ribbon along the central line, dividing the ribbon in two equal parts. The number of twists depends on the shape of the resulting object and is a multiple of this shape parameter *m*.

In Fig. 2 also the *link group* is given, both for the individual resulting elements and for the total object. Depending on the parameters *m* and *n*, complicated graphs result as in Fig. 3a, which are intertwined individual *GML* bodies. The graphical display with closed interlinked bodies can also be shown in a ribbon style with identification (Fig. 3b; the lines or ribbons do not cross in the plane).

Whereas the original *GML* cutting results were based on cutting of full 3D GML bodies and surfaces, with a fixed *GML* and a moving knife, the total possible cuts can also be studied by planar geometry, as suggested by Fig. 2. Figure 4 shows the results of all possible vertex-to-vertex cuts and vertex-to-side cuts for *GML*'s with square and with hexagonal cross section. The results are related to the divisors of *m*. The three rows for the square in Fig. 4 are related to divisors 1, 4 and 2, and the four rows for the hexagon are related to divisors 1, 6, 3 and 2.

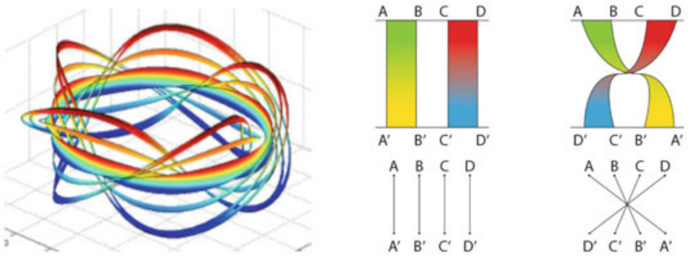


Fig. 3 **a** 3D representation and **b** Ribbon style representation

Lemma 1: *Cutting a fixed GML_m^n body with a moving knife is fully equivalent to rotating the same GML_m^n body through a fixed knife.*

Hence, planar geometry can indeed be used to study the total number of possible ways of cutting and give indications about the shape and link number of the resulting bodies.

Remark 1: This fixed knife can have one or more blades and the number of blades is determined by the divisors of m and thus by the twisting number n. For example in the hexagon in Fig. 4, the second row is obtained using a fixed knife with six blades, and the third row is obtained with a three-bladed knife [1, 11].

Remark 2: Using a fixed knife the *GML* body has only to be rotated over 360° to obtain the full cutting. A moving knife has to perform (n.360°) rotations [1, 11].

Remark 3: Originally the inspiration of reversing to a fixed knife was by how bamboo culms are cut, but it is noted that in his seminal paper on Special Relativity Theory, Einstein makes reference to the well-known fact that Maxwell's laws of magnetism ensure that it makes no difference whether a magnet moves near a fixed conductor, or the magnet is fixed and the conductor is moving [12] (There is an intimate relationship between SRT and Gielis Transformations [13]).

Remark 4: This cutting of polygons is reminiscent of similar problems of cutting polygons with diagonals (which are *VV*-cuts), due to Euler and Cayley [1], but the case of *GML* is more general since also vertex to side or side to side cuts are taken into account. It is also related to dividing a circle with equally spaced points, but then each edge or side has to be counted as well as a *VV*, namely a cut from one point V_i to the next point V_{i+1} on the circle. In cutting polygons, the shortest *VV* cut is from V_i to V_{i+2}.

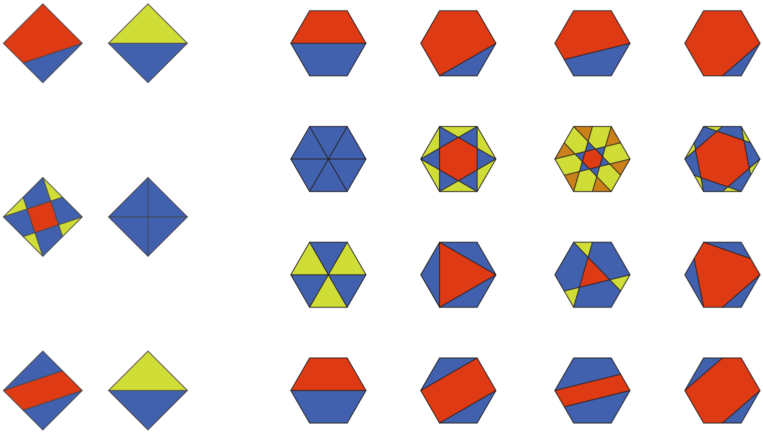


Fig. 4 *VV* and *VS* cuts for GML_4^{4n} and GML_6^n [1]

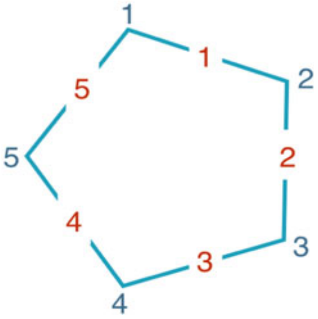


Fig. 5 Numbering of pentagon, vertices in blue, sides in red

3 Cutting Regular Polygons

3.1 The Basic Rules

Consider a regular m-polygon. Vertices and sides are numbered from 1 to m. In Fig. 5 the example of a pentagon is shown.

A regular m-polygon can be cut in various ways, from vertex to vertex (notation VV_{ij}, e.g. $VV_{1,3}$ from vertex 1 to vertex 3), from vertex to side (notation VS_{ik}, e.g. $VS_{1,3}$ from vertex 1 to side 3), or from side to side (notation SS_{kl}, e.g. $SS_{1,3}$ from side 1 to side 3). VV_{max}, VS_{max} and SS_{max} are the cuts from and to vertices or sides with maximal separation (*max* is not necessarily the longest length). For example in Fig. 5, $VV_{max} = VV_{1,3} = VV_{1,4}$; $VS_{max} = VS_{1,3}$, and $SS_{max} = SS_{1,3}$. If the line in SS_{max} cuts side 3 in the middle, and the cut of side 1 is moved to vertex 1, then SS_{max} can be made arbitrarily close to $VS_{max} = VS_{1,3}$.

The following general facts can be observed in Fig. 6:

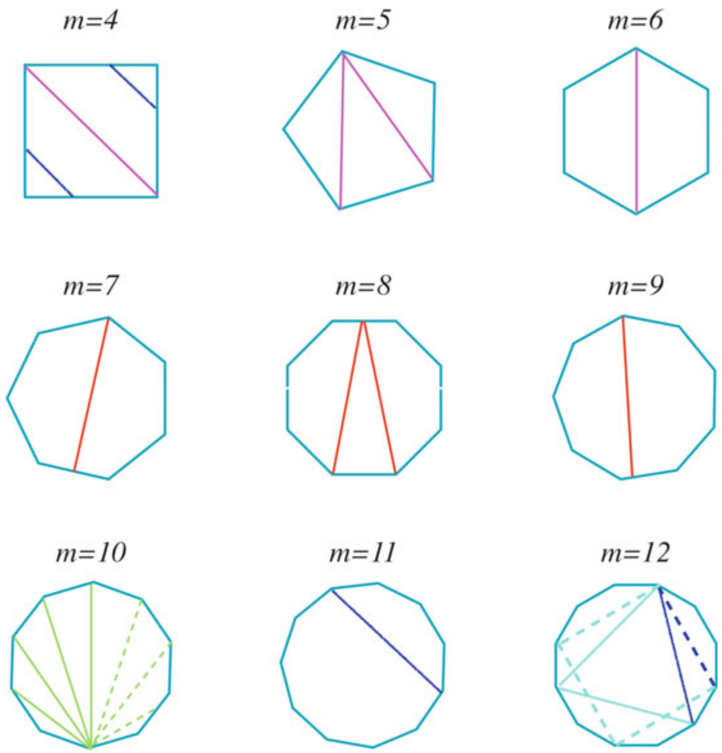


Fig. 6 Understanding modes of cutting

1. In a polygon for even m, VV_{max} goes through the centre of the polygon ($m = 6$), and in a polygon for odd m, VS_{max} goes through the centre, crossing the side opposite the vertex in the middle ($m = 7$).
2. Going from a regular m-polygon to a $(m + 1)$-polygon introduces one extra vertex and one extra side. From $m = 4$ to $m = 5$ a line converts into a wedge of $VV_{1,3} = VV_{1,4}$, giving one extra vertex and one extra side. Considering VS_{max} a wedge is created in m = even polygons ($m = 8$), and one line (through the centre) in m = odd polygons ($m = 7,\ m = 9$).
3. In $m = 10$, all cuts or diagonals are drawn from one vertex. Besides the VV_{max} the other *VV* cuts are two by two symmetrical (solid and dashed lines). In the case of m = odd, the same can be said for *VS* cuts ($m = 7$).
4. A single cut divides an m-polygon into two parts, which are defined by their shape and number of vertices and sides. In $m = 4$ the square is divided by the red diagonal into two triangles. In $m = 6$ the hexagon is divided into two equal quadrilaterals or trapezoids. In $m = 11$ the polygons is divided into an octagonal shape with 8 vertices, and a pentagonal shape with 5 vertices.

3.2 Cutting and Divisors of M

There is a definite relation between the ways of cutting and the number of divisors of m. In Fig. 6, $m = 11$, a $VV_{1,5}$ cut is related to the smallest divisor of m, d_{min}=1. For $m = 12$ in Fig. 6 we have:

1. $VV_{1,4}$ cut (dark blue dashed line), is repeated as $VV_{4,7}$ and as $VV_{7,10}$ and $VV_{10,1}$. In fact, the latter cuts are obtained from the original $V_{1,4}$ cut (blue dashed line) by a rotation by $\frac{2\pi}{4}$. This gives a square, related to divisor 4.
2. $VV_{1,5}$ cut (blue solid line), is repeated as $VV_{5,9}$ and as $VV_{9,1}$. In fact, the latter cuts are obtained from the original $V_{1,5}$ cut (blue solid line) by a rotation by $\frac{2\pi}{3}$. This gives a triangle, related to divisor 3.
3. $VV_{1,3}$ cut is repeated 6 times over an angle of $\frac{2\pi}{6}$, related to divisor 6. This will give a hexagon.
4. $VV_{1,2}$ cut does not divide the polygon, but coincides with the original side 1. Rotation this by $\frac{2\pi}{12}$ will give the original dodecagon. An inscribed dodecagon is obtained via a $SS_{1,2}$ cut, rotated twelve times by $\frac{2\pi}{12}$. Both cases are related to $d_{max} = 12$.

In $m = 4$ the upper blue line is a $SS_{1,2}$ cut from side 1 to side 2. When rotated by 180° the lower blue line is obtained, for the second smallest divisor $d_2 = 2$.

The inscribed figures (triangle and square in $m = 12$) close in one rotation but this can be generalized, for $m = \frac{p}{q}$ with p, q rational numbers and relative prime.

1. In $m = 5$ the sequence $VV_{1,3}$, $VV_{3,5}$, $VV_{5,2}$, $VV_{2,4}$, $VV_{4,1}$ will generate a pentagram in the pentagon, i.e. a figure that closes in 2 rotations, having 5 angles that are spaced $\frac{4\pi}{5} = 144°$ apart. This generates the classic Pythagorean pentagram that led to the discovery of irrational numbers and the golden ratio.
2. In $m = 7$ the sequence $VV_{1,3}$, $VV_{3,5}$, $VV_{5,7}$, $VV_{7,2}$, $VV_{2,4}$, $VV_{4,6}$, $\mathrm{V}V_{6,1}$ generates a heptagram, closing in 2 rotations, corresponding to $m = \frac{p}{q} = \frac{7}{2.}$
3. Also in $m = 7$ the sequence $VV_{1,4}$, $VV_{4,7}$, $VV_{7,3}$, $VV_{3,6}$, $VV_{6,2}$, $VV_{2,5}$, $VV_{5,1}$ generates a heptagram, closing in 3 rotations, corresponding to $m = \frac{p}{q} = \frac{7}{3}$.

3.3 Rotations and Scaling Straight Knives

With these rules the analytic definition of d-knife is a construction, with i straight lines, is:

$$\sin\left(\alpha + \frac{2\pi}{m}i\right)x_i + \cos\left(\alpha + \frac{2\pi}{m}i\right)y_i + \delta = 0 \ , \quad i = 0,\ 1,\ ..,m-1; \quad -\frac{\pi}{m} \leq \alpha \leq \frac{\pi}{m}, \tag{1}$$

The number of straight lines depends on the divisor of m. For any integer m the divisors are numbered from 1 (smallest divisor, knife d_1) to m. The latter is the

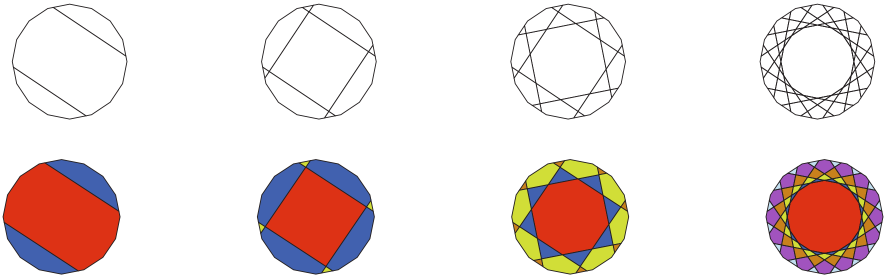
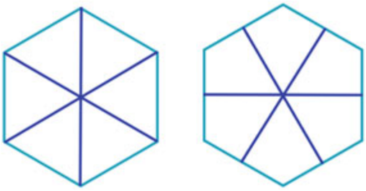


Fig. 7 Cutting with d$_i$ knives (d$_2$ = 2, d$_3$ = 4, d$_4$ = 8, d$_5$ = 16) and the resulting sectors

Fig. 8 Converting VV_{max} into SS cut through centre of the sides by rotation



maximal divisor and is denoted as d$_{max=m}$. For $m = 16$ with 5 divisors we have the following d$_i$ knives: ($d_1 = 1$; $d_2 = 2$; $d_3 = 4$; $d_4 = 8$; $d_{max} = 16$).

A d$_1$ knife is a single line cutting the polygon. Other knifes can be constructed by rotating and translating this knife. For $m = 16$ results of cutting with d$_i$ knives for divisors 2, 4, 8 and 16 are shown in Fig. 7. They are rotations of the d$_1$ knife given by the parameter α in (1). Parameter δ in (1) is called *zooming* parameter since, for example in Fig. 4 rows 2 and 3, the red hexagon and red triangle, respectively can be made larger or smaller, depending on the position of the knives.

The combination of the rotation parameter α and the zooming parameter δ in (1) gives more possibilities. A translation of the knife to a position parallel to the original knife is given by the parameter δ in (1) in the case of a *SS* cut, moving the cutting position along the edge or side. In the case of a *VS* cut, a rotation of the knife with vertex *V* as centre of rotation can be done, to cut the edge or side at another position.

The relation of divisors and rotations show that *VV* and *SS* cuts can be transformed into each other by rotations (Fig. 8):

1. In $m = 6$, the $VV_{max} = VV_{1,4}$ cut or diagonal can be rotated every 60° and all diagonals meet in the centre (Result is 6 diagonals that coincide 2 by 2)
2. This shape can then be rotated by 30°, resulting in SS_{max}. The rotation can in fact be done for any angle.

There are many ways in which the figures can be transformed into the other figures, using rotation and scaling. A first example sequence of rotations and scaling could be:

1. If the red square in Fig. 9**g** is rotated by 45° the result is **c,**
2. If the red square in **c** is scaled to size zero, d results.

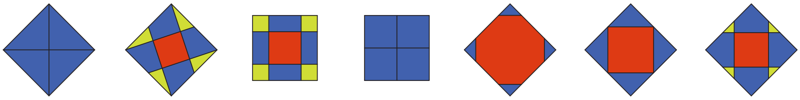


Fig. 9 Ways of cutting square

3. If the cross in **d** is rotated by 45° the result is **a.**

A second example in Fig. 9:

1. If the red square in **f** is scaled to a larger size, **e** results; when it is scaled to **a** smaller size **g** results.
2. When the inscribed figure in **g** is rotated so that one of the sides of the small yellow triangles ends in a vertex, we obtain **b**.

This shows that all figures (for GML_4 with a square cross section) can be considered as transformations of an inscribed square relative to the circumscribed one. For other symmetries this is the relation of the inscribed m-polygon inside an circumscribing m-polygon.

3.4 The Way of Cutting Determines the Final Result

In Fig. 9 from left to right, all possible cuts are shown for a square

1. **a.** One case of *VV* cut, with two VV_{max} diagonals
2. **b.** One case of *VS* cut, with a $VS_{1,2}$ cut and its rotations over $\frac{2\pi}{4}$. It total 3 different shapes are created. Four green triangles, four blue quadrilaterals and one red square. The smaller red square can be considered as the inscribed square rotated and scaled to smaller size.
3. **c** and **d.** Two cases of $SS_{max} = SS_{1,3}$ cuts and rotations. One *SS* cut does not pass through the centre (**c.**) and the other one passes through the centre (**d**). The former creates 3 different sets of quadrilaterals indicated by different colours. The latter creates four different squares. (In *GML* in **d.** these four different squares form one body (compare Fig. 2b, case BII), and in **c.** each of the coloured zones creates 3 separated bodies)
4. Three cases of $SS_{1,2}$ cuts (**e**, **f**, **g**). It is clear that the result depends on where the cut is made. The middle figure **f.** is the inscribed square, while **e.** and **g.** are scaled version (larger and smaller, in this case without rotation).
5. This also generates different shapes. In **e.** four triangles and one octagon; in **f.** four triangles and one square and in **g.** one set of 4 triangles, one set of 4 pentagons and a central square. Again in *GML* and in rational Gielis curves RGC they will form different bodies or layers.

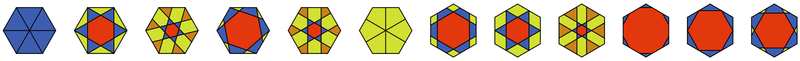


Fig. 10 Hexagon cutting

Note that in Fig. 9 we have 1 *VV*-cut (**a**), 1 *VS*-cut (**b**), and 5 *SS*-cuts, 2 for $SS_{1,3}$ (**c**,**d**) and 3 for $SS_{1,2}$ (**e**,**f**,**g**). The same logic is applied to hexagons in Fig. 10, with more possibilities of cutting, namely two *VV* cuts, two *VS* cuts and eight *SS* cuts. Here we have 2 *VV*-cuts, 2 *VS*-cuts and 8 SS-cuts ($SS_{1,2}$, $SS_{1,3}$ and $SS_{1,4}$).

3.5 Inheritance of Possible Cuts for Different Divisors

In Fig. 11 *VV* and *VS* cuts are shown for $m = 4,\ 6$. A square has three divisors, so 3 rows; for the hexagon there are four rows corresponding to 4 divisors. For *m* prime only 2 rows result, for divisors *m* and 1.

- For even *m* the vertical columns of square and hexagon in Fig. 11 show that the possible cuts are the same for all divisors, as the result of rotating cut 1 over the relevant angle, related to divisor. The identification of vertices and of the knife (e.g. for d_2 =2 lower row Fig. 11) links the cutting of m-polygons to cutting of GML_m^n bodies
- For odd *m* (in case of pentagon in Fig. 12), the *VV* cuts are inherited from cut 1 via rotations. This is not the case for the *VS* cuts of the pentagon in Fig. 12. However, one figure is missing in the upper row, namely the *VS* cut not through the centre. If this figure is also considered, then also for the pentagon the number of cuts is fully inherited for the two divisors, as in the case of the even *m.*
- The reason why in Fig. 12 for the pentagon (and in general for any odd *m*) only one figure is shown for *VS* (the cut through the centre) is that the number of sectors and the number of vertices and sides of the resulting polygons remains the same. In the case of the pentagon and $VS_{1,3} = VS_{max}$, two quadrilateral figures are created, whether or not the cut goes through the centre. The quadrilateral shapes share the topological characteristic of four vertices and four sides.
- In case of the $VS_{1,3} = VS_{max}$ cut going through the centre, the two quadrilateral shapes have also the exactly same shape, a geometrical characteristic. If $VS_{1,3}$ does not pass through the centre, then the geometrical shapes of the two quadrilaterals are different.

So, if this **geometrical** characteristic is considered, also for divisor 1 or one cut, two different shapes need to be considered, whereby the two shapes are different or the same. In the **topological** case, these two shapes reduce to one shape as in Fig. 12. The same line of reasoning can be considered for *SS* cuts, where in Fig. 12 there are 2 line cuts in upper row versus 8 in the lower row.

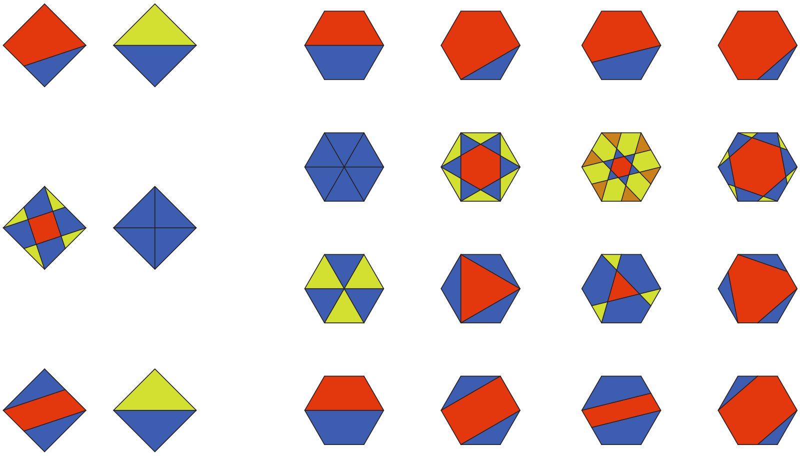


Fig. 11 VV and VS cuts of square and hexagon

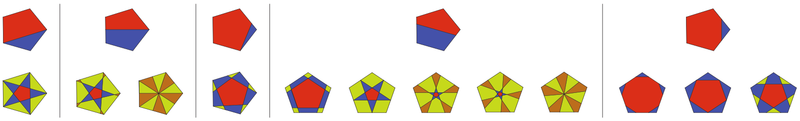


Fig. 12 All possible cuttings of a pentagon via VV, VS and SS cuts

In general, starting from the maximum divisor, or max cuts equal to m, for other divisors the number of possibilities is inherited precisely, in the **geometrical** sense. This leads to Theorem 1.

4 The Geometrical Solution

Theorem 1: *The total number of different ways of cutting an m-polygon* Ξ_m^{geo} *is the number of 1 or m cuts, times the number of divisors of m.*

- *For even m* $(= 2k)$: $\Xi_m^{geo} = N_m^{div}\left(m + 1 + N_{m-2}^{SS}\right)$
- *For odd m* $(= 2k + 1)$: $\Xi_m^{geo} = N_m^{div}\left(m + 2 + N_{m-2}^{SS}\right)$

Proof: The total number of ways of cutting an m-polygon according to the rules described above with d-knives for the geometrical case, for m = even and m= odd respectively is as follows:*VV, VS* and *SS* cuts increase from by 1, 1 and 3 respectively from a given even or odd number to the next even or odd number. As a result, the total number of ways of cutting using a d_m knife increases by 5 to each subsequent even or odd number (Table 1, Subtotal), which gives the sequence (2), 2, 7, 7, 12, 12, 22, 22, 27, 27, 32, 32, 37, 37 … for m = 2, 3, 4, 5…15. Taking sums the sequence 4, 14, 24, 34, 44, 54, 64, 74… results, which is monotonically increasing.

Table 1 Number of possible cuts for even and odd m

m = even	Cut type	2	4	6	8	10	12	14
(m − 2)/2	VV	0	1	2	3	4	5	6
(m − 2)/2	VS	0	1	2	3	4	5	6
Step +3	SS	2	5	8	11	14	17	20
	Subtotal	2	7	12	17	22	27	32
	Divisors	2	3	4	4	4	6	4
Total		**4**	**21**	**48**	**68**	**88**	**162**	**128**
m = odd	Cut type	3	5	7	9	11	13	15
(m − 3)/2	VV	0	1	2	3	4	5	6
(m + 1)/2	VS	2	3	4	5	6	7	8
Step + 3	SS	5	8	11	14	17	20	23
	Subtotal	7	12	17	22	27	32	37
	Divisors	2	2	2	3	2	2	4
Total		**14**	**24**	**34**	**66**	**54**	**64**	**148**

Since the number of possibilities is determined by d_1 and d_m and is inherited by the other divisors given identification of vertices and knives (Fig. 13 for divisors 1, 2 and 6 in a hexagon), the total number is then the subtotal times the number of divisors. The identification links planar geometry to 3D GML_m^n bodies. If one follows the d_1-knife along the basic line of the GML_6^6 body, the different positions of the knives indicated by arrows in Fig. 13 for d_1 function as a clock, relative to the torus circumscribing the GML_6^6 body. For other knives the clock arithmetic is the same, albeit with more hands.

Table 2 gives the product of the subtotals with positive integers. Entries in rows can be computed as $u_n = u_{n-1} + u_{n-2} - u_{n-3}$, with $u_0 = 2$ for even and $u_0 = 3$ for odd numbers. In bold red are the totals of Table 1 for m odd and in bold green for m even.

The number of possible cuts can be given by a recurrence formula. For N_m, the number of ways of cutting an m-gon for one divisor, whereby N_m^{SS} stands for the number of SS cuts for m and N_{m-2}^{SS} for the number of SS cuts for the polygon with $(m - 2)$ (i.e. the previous odd or even number) and with k a natural number.

- For even m $(= 2k)$: $N_{m=2k} = m + 1 + N_{m-2}^{SS}$ (1a)
- For odd m $(= 2k + 1)$: $N_{m=2k+1} = m + 2 + N_{m-2}^{SS}$ (1b)

If the number of SS cuts is kept separate, taking into account the step +3, this part of the general formula is recursive. Since $N_m^{SS} = \left(N_{m-2}^{SS} + 3\right)$, it follows that $N_{m-4}^{SS} + 6 = N_{m-6}^{SS} + 9 = \ldots$

Because of the exact inheritance for the geometrical case the total number of ways of cutting for all divisors is then the above formula times the number of divisors N_m^{div} of a number m.

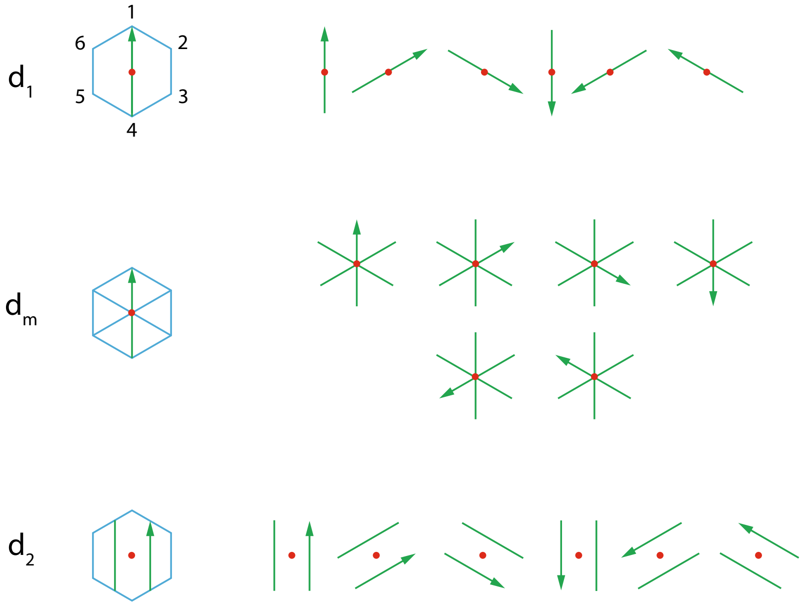


Fig. 13 Inheritance from d_1 by all d-knives through identification

Table 2 Product of subtotals of Table 1 and number of divisors

N° divisors	2	7	12	17	22	27	32	37
2	**4**	**14**	**24**	**34**	44	**54**	**64**	74
3	6	**21**	36	51	**66**	81	96	111
4	8	28	**48**	**68**	**88**	108	**128**	**148**
5	10	35	60	85	110	135	160	185
6	12	42	72	102	132	**162**	192	222

Remark 5: When considering polygons with convex sides**,** $VV_{i,i+1}$ are possible, so the number of cuts increases by m. When knives are used to cut a circle from equally spaced points, the $VV_{i,i+1}$ cuts need to be added, in particular m cuts.

- For even m $(= 2k)$: $N_{m=2k} = 2m + 1 + N^{SS}_{m-2}$ (2a)
- For odd m $(= 2k + 1)$: $N_{m=2k+1} = 2m + 2 + N^{SS}_{m-2}$ (2b)

References

1. Gielis, J., Tavkhelidze, I.: The general case of cutting of GML surfaces and bodies (2019). https://export.arxiv.org/ftp/arxiv/papers/1904/1904.01414.pdf

2. Tavkhelidze, I.: On the some properties of one class of geometrical figures and lines. In: Reports of Enlarged Sessions of the Seminar of I. Vekua Institute of Applied Mathematics vol. 16, no.1. pp. 35–38 (2001)
3. Gielis, J.: A generic geometric transformation that unifies a wide range of natural and abstract shapes. Am. J. Bot. **90**(3), 333–338 (2003)
4. Gielis, J., Haesen, S., Verstraelen, L.: Universal natural shapes - from the supereggs of Piet Hein to the cosmic egg of Georges Lemaître. Kragujevac J. Math. **28** (2005)
5. Guitart, R.: Les coordonnées curvilignes de Gabriel Lamé–Représentation des situations physiques et nouveaux objets mathématiques. Bulletin de la Sabix. Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique (44), 119–129 (2009)
6. Gielis, J., Tavkhelidze, I., Ricci, P.E.: About, "bulky" links generated by generalized Möbius-Listing bodies. J. Math. Sci. **193**(3), 449–460 (2013)
7. Tavkhelidze, I., Cassisa, C., Gielis, J., Ricci, P.E.: About bulky links, generated by generalized Mobius-listing's bodies GML_3^n. Rendic. Lincei Mat. Appl. **24**, 11–3 (2013)
8. Tavkhelidze, I., Ricci, P.E.: Some properties of "bulky" links, generated by generalised Möbius–listing's bodies. In: Modeling in Mathematics, pp. 159–185. Atlantis Press, Paris (2017)
9. Pinelas, S., Tavkhelidze, I.: Analytic representation of generalized Möbius-listing's bodies and classification of links appearing after their cut. In: International Conference on Differential & Difference Equations and Applications, pp. 477–493. Springer, Cham (2017)
10. Tavkhelidze, I.: About connection of the generalized Möbius listing's surfaces with sets of ribbon knots and links. In: Proceedings of Ukrainian Mathematical Congress S. 2, Topology and Geometry - Kiev, 2011, pp. 117–129 (2011)
11. Tavkhelidze, I., Gielis, J.: The process of cutting GML_m^n bodies with d_m knives. Reports of the Enlarged Sessions of the Seminar of I. Vekua Institute of Applied Mathematics, vol. 32 (2018)
12. Einstein, A.: On the electrodynamics of moving bodies. Ann. Phys. **17**(891), 50 (1905)
13. Gielis, J.: The Geometrical Beauty of Plants. Atlantis Springer (2019)

Computational Simulation of Bacterial Infections in Surgical Procedures: An Exploratory Study

J. A. Ferreira, Paula de Oliveira, and Pascoal M. Silva

Abstract These last years the insertion of implants and medical devices has emerged as a common surgical procedure. Following their implantation, the bacteria, inoculated during the surgery, or coming from a preexisting focus of infection, can colonize a significant proportion of them. The resistance of bacteria against antibiotics increases dramatically once a colony forms. Researchers of different fields are working on the development of new strategies to destroy such colonies: the dispersion of antibacterial drugs in the coatings, or the use of antiadherent coatings are common approaches. The first issue is addressed in this paper. A mathematical model of sustained drug delivery, from a medical implant, and its action on the bacterial fight, is presented. The model is composed by coupled systems of Partial Differential Equations that describe the release of drug and the evolution of the biotic population. The fate of the bacterial infection is analyzed as a function of the initial contamination during surgery, the permeability of the coating and the dissolution rate of the drug. Computational simulations will give a lively picture of the process.

Keywords Computational simulation · Bacterial infections · Partial differential equations

1 Introduction

Each year millions of medical devices are implanted through common surgical procedures worldwide. More and more patients receive orthopedic implants,

J. A. Ferreira (✉) · P. de Oliveira
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: ferreira@mat.uc.pt

P. de Oliveira
e-mail: poliveir@mat.uc.pt

P. M. Silva
Instituto Politécnico de Coimbra, ISEC, DFM, Rua Pedro Nunes, 3030-199 Coimbra, Portugal
e-mail: pascals@isec.pt

CMUC, University of Coimbra, Coimbra, Portugal

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_32

coronary stents, catheters, pacemakers, valves, cochlear implants, breast implants, dental implants, and intraocular or therapeutic contact lenses. However, a significant proportion of each type of these devices can be colonized by bacteria and becomes the focus of implant-related infections. This fact can be explained by the inoculation of bacteria during the surgery, the existence of a focus of infection in the patient or by the simultaneous action of these two causes. The more septic is the operating theatre and the longer the surgery takes, the more chance there is for bacteria to breed. The species *Staphylococcus aureus* and *Staphylococcus epidermidis* are two of the most common bacteria related to infections of implanted medical devices [5, 12].

It is expected that the increasing use of implantable devices will lead to a rise in the number of related infections. This happens due to two main reasons. The first one is the damagement of epithelial barriers during surgery, and the consequent impairment of host defense mechanisms. Bacteria from the patient's skin or mucous can contaminate the device during the implantation. Bacteria can also come from the hands of the staff, or from the hospital environment. The second reason is bacterial adherence. When bacteria adhere to surfaces, they have a larger likelihood to survive because nutrients, which are in suspension in the surrounding fluids, deposit on the surfaces, leading to an increase of their local concentration. Most strains of those bacteria form then biofilms, which are clusters of cells embedded in a matrix. The resistance of bacteria against the host immune system and antibacterial agents increases dramatically once a biofilm forms. Accordingly, it is essential to eradicate bacteria in the first hours following the surgery [6, 8].

The most common approaches for preventing the biofilm formation (Fig. 1) and avoiding infections are (i) the dispersion of antimicrobial agents in a polymeric coating of the implant, and/or (ii) the use of anti-adherent coatings [3, 7, 13]. Modelling and computational simulation of those strategies, complemented by the simultaneous prediction of bacterial densities can represent an important coadjutant to laboratorial experiments. In this paper we will address issue (i), from a mathematical point of view. We present a mathematical model to predict the delivery of drug from a biodegradable coating of a medical implant and, simultaneously, to predict the evolution of a bacterial population. The model is composed by a set of coupled systems

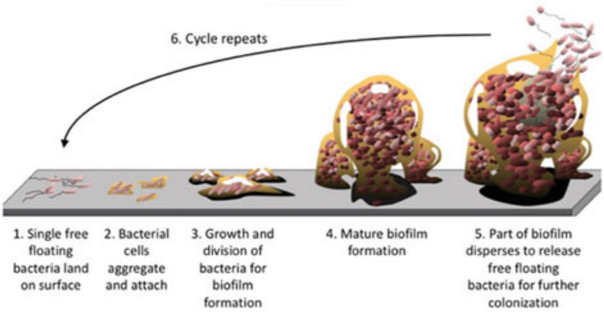

6. Cycle repeats
1. Single free floating bacteria land on surface
2. Bacterial cells aggregate and attach
3. Growth and division of bacteria for biofilm formation
4. Mature biofilm formation
5. Part of biofilm disperses to release free floating bacteria for further colonization

Fig. 1 Biofilm formation - Adapted from *www.zmescience.com/science/what-are-biofilms*

of Partial Differential Equations, that represent the evolution of an antibacterial agent and its action against the colony of bacteria inoculated during surgery. We assume that these bacteria are homogeneously dispersed in the surface of the implant, in the moment of the insertion. Assuming this initial form of sepsis, we are aware that a huge number of factors influence the fate of bacterial fighting.

In this exploratory study we analyze the influence of the initial bacterial contamination and of the properties of the polymer coating and the drug on the bacterial behavior. Regarding the polymer properties we illustrate the effect of the coating permeability; in what concerns the drug we simulate the dependence on its dissolution rate. At the best of our knowledge, the model represents an original contribution as it simulates an *in vivo* interplay of some of the main actors of the process. We remark that the mathematical model studied here is an improvement of the one considered in [1] where the drug delivery mathematical model was coupled with a one dimensional model for the bacterial density. In Sect. 2 we present the mathematical model. In Sect. 3 we exhibit a set of numerical simulations to illustrate the behavior of the model. Namely, in Sect. 3.1 we present simulations of the global behavior of the concentrations of the solid drug, the dissolved drug and the interstitial fluid. The evolution of the bacterial density is also analyzed. In Sect. 3.2 we illustrate the influence of the initial contamination in the behavior of the bacteria. The influence of the coating permeability and the drug dissolution coefficient are also exhibited. Finally in Sect. 4 some conclusions are raised.

2 Mathematical Model

In what follows we use the following notation. Let Ω be a two-dimensional open domain and $[0, T]$ a time interval. If $u : \overline{\Omega} \times [0, T] \to \mathrm{I\!R}$ is a function then, for $t \in [0, T]$, by $u(t)$ we represent the function $u(t) : \overline{\Omega} \to \mathrm{I\!R}$ given by $u(t)(x) = u(x, t)$, $x \in \overline{\Omega}$. We represent in Fig. 2 the geometry of the model: $\overline{\Omega}_1$ stands for a biodegradable polymeric coating of a metallic implant and $\overline{\Omega}_2$ represents the adjacent tissue. The drug is initially dispersed in Ω_1 in the solid state. When it enters in contact with the interstitial fluid, that permeates the surrounding tissue Ω_2, it dissolves progressively and the drug is delivered through the interface $\partial\Omega_{1,2}$. The boundary $\partial\Omega_{1,left}$ represents the interface between the polymeric coating and a metal implant. We assume that there are no fluxes – of interstitial fluid, drug or bacteria - through this boundary. An initial concentration of bacteria, resulting from the inoculation during the surgery, is considered on the interface $\partial\Omega_{1,2}$.

The unknowns of the model are the concentration of interstitial fluid c_ℓ, the concentration of the solid drug c_s, the dissolved drug c_d and the density of the bacterial population c_b.

The cascade of phenomena that occurs is described by the permeation of the interstitial fluid in the porous biodegradable coating Ω_1, the dissolution of the solid drug in Ω_1, the diffusion of the dissolved drug through Ω_1 and Ω_2 and the fight against the bacterial population.

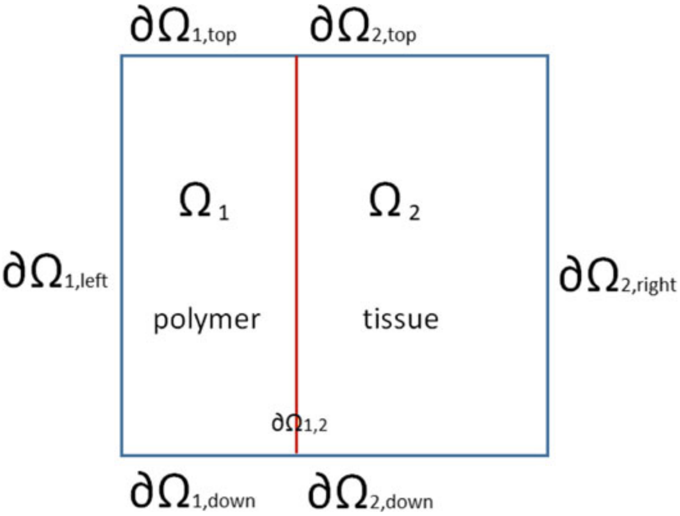


Fig. 2 Spatial domain

1. Polymeric coating – domain Ω_1
   The behavior of the concentrations of the interstitial fluid, c_ℓ, the solid drug concentration, c_s, and the dissolved drug, c_{d1}, in Ω_1, are governed by the following equations

$$\begin{cases} \dfrac{\partial c_\ell}{\partial t}(t) = \nabla.(D_\ell(t)\nabla c_\ell(t)) \\ \\ \dfrac{\partial c_{d1}}{\partial t}(t) = \nabla.(D_{ef}(t)\nabla c_{d1}(t)) + f(c_s(t), c_{d1}(t), c_\ell(t)) - R_{db}c_{d1}(t)c_b(t) \\ \\ \dfrac{\partial c_s}{\partial t}(t) = -f(c_s(t), c_{d1}(t), c_\ell(t)) \end{cases}, \tag{1}$$

for $t \in (0, T]$. In (1), D_ℓ represents the diffusion coefficient of the interstitial fluid in the polymeric coating. We consider that Ω_1 is a biodegradable porous medium and that D_ℓ dependents on time. Accordingly the diffusion coefficient D_{ef} of the dissolved drug is also time dependent. For the time evolution of the porosity $\epsilon(t)$ due to the polymeric coating degradation we take

$$\epsilon(t) = \epsilon_0 + (1 - \epsilon_0)(1 + e^{-2k_d t} - e^{-k_d t})$$

that was introduced in [11]. In this last expression ϵ_0 stands for the initial porosity of the polymeric coating and k_d represents the degradation rate. The diffusion coefficient of the interstitial fluid is represented by

$$D_\ell(t) = (\epsilon(t))^{\frac{3}{2}} D_{\ell,0},$$

where $D_{\ell,0}$ represents the initial diffusion in the non degraded coating [9]. The diffusion coefficient of the dissolved drug is a weighted mean

$$D_{ef}(t) = \frac{1 - \epsilon(t))D_1 + \widetilde{k}\epsilon(t)D_2}{1 - \epsilon(t) + \widetilde{k}\epsilon(t)},$$

where D_1 stands for the drug diffusion coefficient in the solid part of the polymer, D_2 represents the drug diffusion in the polymer pores filled with fluid [11]. We note that $\widetilde{k}$ denotes the drug partition coefficient, between the liquid filled pores and the solid polymer. For simplicity we take $\widetilde{k} = 1$. The consumption of drug by the bacterial population is represented by the term $R_{db}c_{d1}(t)c_b(t)$, where R_{db} stands for a positive constant. The reaction term f represents the rate of conversion of solid drug into dissolved drug and is defined by

$$f(c_s(t), c_{d1}(t), c_\ell(t)) = \alpha H(c_s(t)) \frac{c_{sol} - c_{d1}(t)}{c_{sol}} c_\ell(t),$$

where α is the dissolution rate, H is the Heaviside function and c_{sol} represents the solubility limit concentration [10].

2. Adjacent tissue – domain Ω_2
The evolution of the dissolved drug concentration in Ω_2, c_{d2}, is described by

$$\frac{\partial c_{d2}}{\partial t}(t) = \nabla.(D_{d2}\nabla c_{d2}(t)) - R_{db}c_{d2}(t)c_b(t), \tag{2}$$

for $t \in (0, T]$, where D_{d2} represents the diffusion coefficient. This coefficient is space dependent due to the fact that bacteria have been inoculated during the surgery and consequently diffusion assumes different values in $\Omega_{2,b}$ and $\Omega_{2,nob} = \Omega_2 \backslash \Omega_{2,b}$, where $\Omega_{2,b}$ represents the spatial domain occupied by bacteria at initial time. We define

$$D_{d2}(x) = \begin{cases} Tol, & x \in \Omega_{2,b} \\ \\ D_{d2nob}, & x \in \Omega_{2,nob} \end{cases},$$

with $Tol < D_{d2nob}$.

3. Polymeric coating and adjacent tissue – domain $\Omega_1 \cup \Omega_2$
The density of bacteria, c_b, is governed by

$$\frac{\partial c_b}{\partial t}(t) = \nabla.(D_{bact}\nabla c_b(t)) + Fun(c_d(t))c_b(t), \tag{3}$$

for $t \in (0, T]$, where $c_d = c_{d1}$ in Ω_1 and $c_d = c_{d2}$ in Ω_2, the diffusion coefficient D_{bact} depends on space and is defined by

$$D_{bact}(x) = \begin{cases} D_{b1}, & x \in \Omega_1 \\ \\ D_{b2}, & x \in \Omega_2 \end{cases}.$$

The net proliferation of the bacteria is defined by

$$Fun(c) = \lambda - \frac{E_{max} c^{\gamma}}{c_{50}^{\gamma} + c^{\gamma}}. \tag{4}$$

Equation (4) represents the balance between proliferation and the antibacterial action of the drug. This action is defined by Hill model, that is extensively used in the literature. We believe that one of the reasons for its success is its flexibility and effectiveness in fitting experimental data [4]. It includes the two main pharmacodynamic properties of a drug: the maximum effect (E_{max}) and the concentration producing 50% of the maximum effect (c_{50}). More precisely, E_{max} represents the maximum effect which can be expected from the drug: when this magnitude of effect is reached, increasing the dose will not produce a greater magnitude of effect. In Eq. (4), γ represents a measure of the cooperation between bacteria. If $\gamma = 1$ the adhesion of the bacteria to the surfaces is independent of each other. If $\gamma > 1$, then there is cooperation, and if $\gamma < 1$ no cooperation occurs. We will consider $\gamma = 1$.

Coupled systems (1)–(3) are completed with the following initial, boundary and interface conditions:

- Initial conditions:

$$c_{\ell}(0) = c_{d1}(0) = 0,\ c_s(0) = c_{s,i},\ c_b(0) = 0 \text{ in } \Omega_1,\ c_{d2}(0) = 0,\ c_b(0) = c_{b,i} \text{ in } \Omega_2.$$

- Boundary conditions:

  – $\partial\Omega_{1,left}$ is isolated that means that $J_c(t).\eta_1 = 0\, on\ \partial\Omega_{1,left},\ t \in (0, T]$, for $c = c_{\ell}, c_{d1}, c_b$, where $J_c(t)$ denotes the flux of c and η_1 represents the unitary exterior normal to Ω_1,

  – $\partial\Omega_{2,right}$ is isolated that is

$$J_c(t).\eta_2 = 0 \text{ on } \partial\Omega_{2,right},\ t \in (0, T],$$

  for $c = c_b, c_{d2}$, where, as before, $J_c(t)$ denotes the flux of c and η_2 represents the unitary exterior normal to Ω_2,

  – symmetry conditions on $\bigcup\limits_{i=1,2,j=top,down} \partial\Omega_{i,j}$ that are mathematically defined by

$$\frac{\partial c}{\partial x_2}(t) = 0 \text{ on } \bigcup_{j=top,down} \partial\Omega_{1,j}, t \in (0, T],$$

for $c = c_\ell, c_{d2}, c_b$, and

$$\frac{\partial c}{\partial x_2}(t) = 0 \text{ on } \bigcup_{j=top,down} \partial \Omega_{2,j}, t \in (0, T],$$

for $c = c_{d2}, c_b$.

- Interface conditions:
  On the common boundary of Ω_1 and Ω_2, $\partial \Omega_{1,2}$, we assume that the fluid flux is proportional to the difference between the fluid concentration on the boundary and the fluid concentration c_{ext} in Ω_2, that is $J_{c_\ell}(t).\eta_2 = \beta(c_\ell(t) - c_{ext}) on \partial \Omega_{1,2}, t \in (0, T]$, where β is related with the permeability of the interface that, to simplify, we assume time independent. For the dissolved drug concentration we assume the continuity of the concentration and of the flux, that is

$$c_{d,1}(t) = c_{d,2}(t), \;\; J_{c_{d1}}(t).\eta_1 + J_{c_{d2}}(t).\eta_2 = 0 \;\text{ on } \partial \Omega_{1,2}, \; t \in (0, T].$$

3 Numerical Simulations

The problem is solved for the first 7 h after surgery and considering that bacteria have been inoculated during the procedure.

3.1 *Drug Distribution and Bacterial Dynamics*

In this section we begin by exhibiting global pictures of drug distribution (Daptomycin) and of bacterial evolution using the values presented in Table 1.

Table 1 Parameter values used in the numerical simulations

Parameter (unit)	Value	Parameter (unit)	Value
$D_{\ell,0}$ (m^2/s)	10^{-9}	D_1 (m^2/s)	5×10^{-10}
D_2 (m^2/s)	$2D_1$	D_{d2nob} (m^2/s)	$2D_2$
D_{tol} (m^2/s)	$D_{d2nob}/10$	D_{b1} (m^2/s)	5×10^{-11}
D_{b2} (m^2/s)	10^{-11}		
α $(1/s)$	5×10^{-4}	c_{sol} (mol/mm^3)	2
k_d	3×10^{-4}	ϵ_0	5×10^{-2}
$c_{s,i}$ (mol/mm^3)	5	c_{ext} (mol/mm^3)	1
β (m/s)	1×10^{-4}	$L_1, \; L_2$ (mm)	2, 3
R_{db} $(m^3/(mol.s))$	1×10^{-5}		

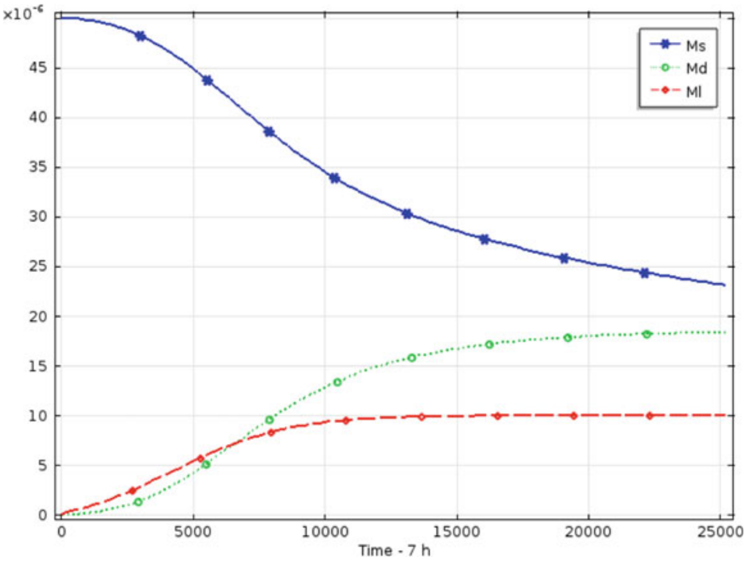


Fig. 3 Behavior of the mass of interstitial fluid $M_\ell(t)$, solid drug $M_s(t)$ and dissolved drug $M_d(t)$ for $t \in [0, 7h]$ in Ω_1

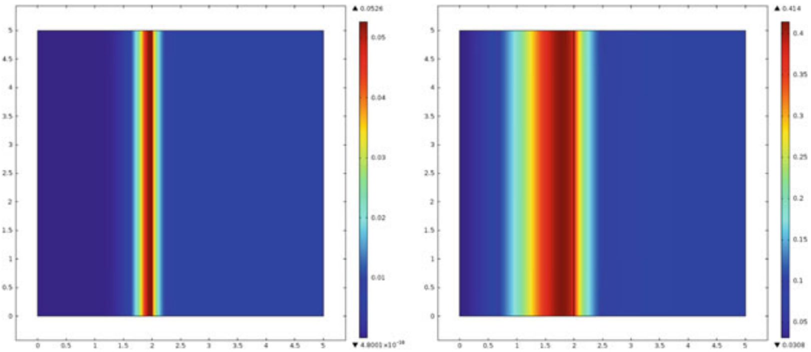


Fig. 4 Dissolved drug distribution at $t = 10$ (min) and $t = 1$ (h)

The following parameters related with Daptomycin are used [2]: $E_{max} = 4\ (h^{-1})$, $c_{50} = 0.5,\ \ \lambda = 0.6\ (h^{-1})$. In Fig. 3 a global picture of the masses of interstitial fluid, solid and dissolved drug, in Ω_1, is exhibited. The mass of interstitial fluid in Ω_1, $M_\ell(t)$, increases over time until a steady state is reached. The mass of solid drug in Ω_1, $M_s(t)$, decreases as the interstitial fluid permeates the polymer and accordingly the dissolved drug mass in Ω_1, $M_d(t)$, increases.

The dissolved drug distribution at $t = 10$ (min) and $t = 1$ (h) is represented in Fig. 4. The permeation of the dissolved drug in the polymer coating and the surrounding tissue is clearly observed.

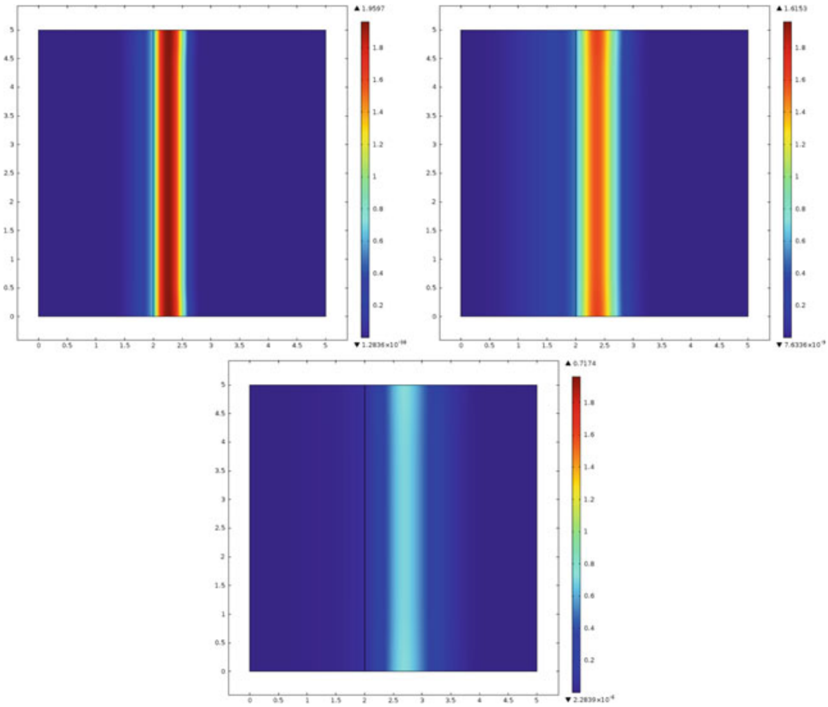


Fig. 5 Bacterial distribution at $t = 10$ m, $t = 1$ h and $t = 3$ h

The distribution of bacteria at $t = 10$ min, $t = 1$ h and $t = 3$ h is shown in Fig. 5. Initially the bacteria inoculated are on the polymer/tissue interface zone. It can be observed that, as time increases, bacterial density decreases. For the set of parameters used in the simulation the drug fights effectively the bacterial population. We note that in Fig. 4 and 5 the scales in the plots are different.

In Fig. 6 we plot the bacteria mass $M_b(t)$. It can be observed that by the first hour of release the bacteria mass stops increasing and it is almost null after 7 h.

3.2 Influence of Parameters

In this section we illustrate the influence of three different parameters: the initial bacterial mass, the permeability coefficient (β) and the dissolution rate of the drug (α).

Non aseptic surgery
We begin by analyzing the effect of surgical contamination by considering that an initial bacterial population enters the patient on the surface of the medical device, that is on $\partial\Omega_{1,2}$.

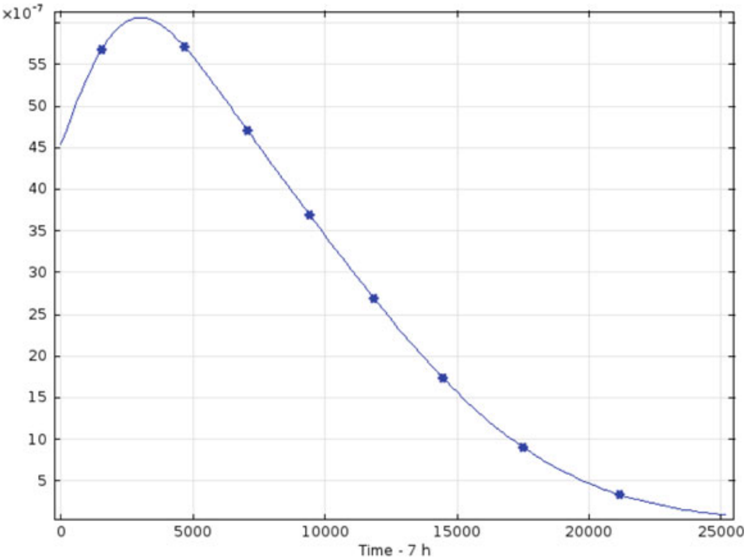


Fig. 6 Evolution of bacteria mass $M_b(t)$ during 7 h

In Fig. 7 the influence of the initial bacterial population on the time evolution of the total mass of bacteria $M_b(t)$ is illustrated for $t \in [0, 7h]$. For an high initial bacterial density ($c_{b,i} = 18,\ 20$) the infection evolves and the drug has no efficacy. Otherwise, for lower a bacterial density ($c_{b,i} = 1.8$), $M_b(t)$ tends to zero. These results illustrate the crucial importance of aseptic conditions in the procedure and in the hospital environment.

Permeability of the coating

The influence of the permeability of the device coating, represented by β, in the evolution of $M_b(t)$ is illustrated in Fig. 8. When β increases, a larger amount of interstitial fluid permeates the coating, more dissolved drug is available and the release is enhanced. Accordingly $M_b(t)$ decreases.

Influence of the dissolution rate of the drug

The influence of the dissolution rate on the evolution of the bacteria mass during 7 h, is illustrated in Fig. 9. As it can be observed, $M_b(t)$ is very sensitive to the dissolution rate. The dissolution coefficient of an antibacterial drug can dictate the fate of a bacterial infection. For $\alpha = 5 \times 10^{-4}$ the infection is not quelled; if the drug has a dissolution coefficient twice that value, then the infection is eliminated.

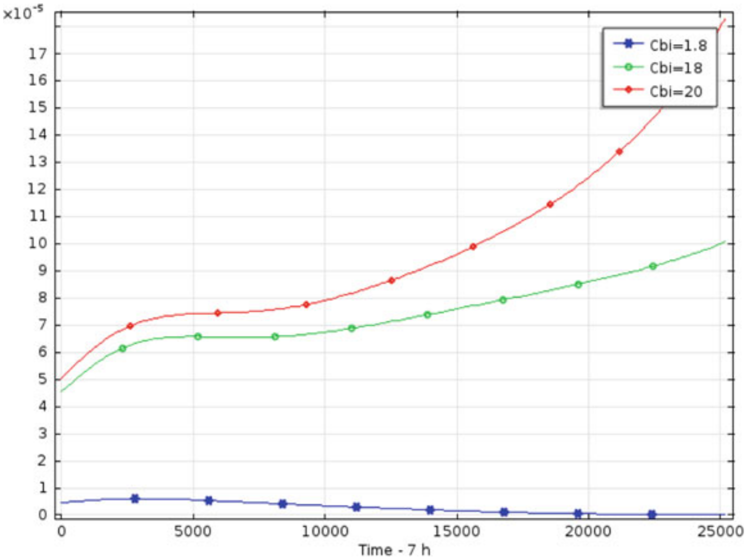


Fig. 7 Evolution of bacterial mass $M_b(t)$ for $c_{b,i} = 1.8, 18, 20$, during 7 h

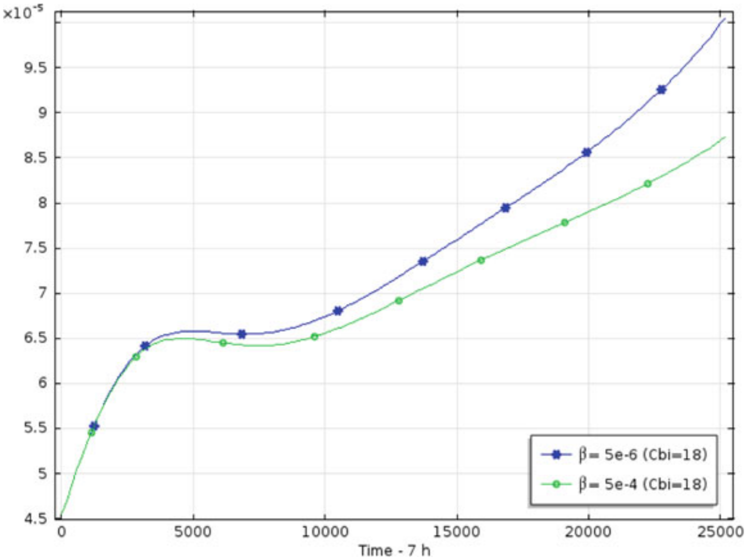


Fig. 8 Evolution of bacterial mass $M_b(t)$ for different coating permeability $\beta = 5 \times 10^{-6}, 5 \times 10^{-4}$, during 7 h

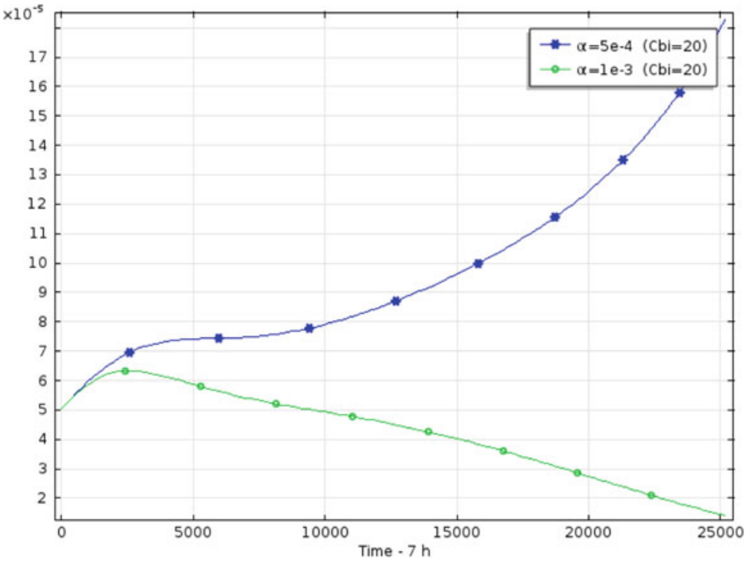


Fig. 9 Evolution of bacterial mass $M_b(t)$ for different dissolution rates $\alpha = 5 \times 10^{-4}$, 10^{-3}, during 7 h

4 Conclusion

When medical devices are inserted through surgical procedures the floating bacteria, inoculated during the surgery, adhere to the foreign surface, and can form a biofilm. This biofilm protects them against the host immune system and the antibacterial drugs. To avoid biofilm formation medical devices can be coated with a polymer layer, where an anti-bacterial drug is dispersed. We present in this paper a mathematical model based on three coupled systems of partial differential equations, that govern the kinetics of an anti-bacterial drug, eluted from a medical implant, and its action on a bacterial population. We carried on several numerical simulations that suggest the following preliminary results:

1. Aseptic procedure and operating theatre: It is commonplace to say that an aseptic operation theatre is crucial to prevent bacterial infections. In Fig. 7 we quantify this assertion. The plots in this figure suggest that if the initial contamination exceeds a certain threshold then it is very difficult to fight the infection.
2. Permeability of the polymeric coating: As suggested by Fig. 8, the permeability of the coating has a meaningful effect on the evolution of the bacterial population. A larger permeability coefficient, β, enhances the permeation of the interstitial fluid that dictates the dissolution of the solid drug. More dissolved drug is released and the likelihood of eliminating the bacterial population increases.

3. Dissolution rate: In Fig. 9, we illustrate the dependence of the bacterial population on the dissolution rate. The results suggest that the drug dissolution rate is a key parameter for controlling the evolution of the bacterial density.

We are aware that the present study has an exploratory character, for several reasons, namely the use of a simplified geometry, the assumption that the initial contamination is homogeneous and the lack of a chemoattractant term in the bacterial equation. We plan to address these problems in the near future.

Acknowledgements This work was supported by Centro de Matemática da Universidade de Coimbra UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MCTES and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020. The first and the second authors were also supported by the project *NEXT.parts – Next- generation of advanced hybrid parts*, funded by EU's Horizon 2020 science programme (Portugal 2020, COMPETE 2020).

References

1. Bernardes, R., Ferreira, J.A., Grassi, M., Nhangumbe, M., de Oliveira, P.: Fighting opportunistic bacteria in drug delivery medical devices. SIAM J. Appl. Math. **79**(6), 2456–2478 (2019)
2. Begic, D., von Eiff, C., Tsuji, B.: Daptomycin pharmacodynamics against Staphylococcus aureus hemB mutants displaying the small colony variant phenotype. J. Antimicrob. Chemother. **63**, 977–981 (2009)
3. Gallo, J., Holinka, M., Moucha, C.: Antibacterial surface treatment for orthopaedic implants. Int. J. Mol. Sci. **15**, 13849–13880 (2014)
4. Gesztelyi, R., Zsuga, J., Kemeny-Beke, A., Varga, B., Juhasz, B., Tosaki, A.: The Hill equation and the origin of quantitative pharmacology. Arch. Hist. Exact Sci. **66**, 427–438 (2012)
5. Gutiérrez, D., Hidalgo-Cantabrana, C., Rodríguez, A., García, P., Ruas-Madiedo, P.: Monitoring in real time the formation and removal of biofilms from clinical related pathogens using an impedance-based technology. PLoS ONE **11**, 0163966 (2016)
6. Rabin, N., Zheng, Y., Opoku-Temeng, C., Du, Y., Bonsu, E., Sintim, H.: Biofilm formation mechanisms and targets for developing antibiofilm agents. Future Med. Chem. **7**, 493512 (2015)
7. Romanò, C., Tsuchiya, H., Morelli, I., Battaglia, A.G., Drago, L.: Infection Antibacterial coating of implants: are we missing something? Bone Joint Res. **8**, 199–206 (2019)
8. Sharma, D., Misba, L., Khan, A.: Antibiotics versus biofilm: an emerging battle ground in microbial communities. Antimicrob. Resist. Infect. Control **8**, 76 (2019)
9. Shen, L., Chen, Z.: Critical reviewof the impact of tortuosity on diffusion. Chem. Eng. Sci. **62**, 3748–3755 (2007)
10. Siepmann, J., Siepmann, F.: Modeling of diffusion controlled drug delivery. J. Controlled Release **161**, 351–362 (2012)
11. Zhu, X., Braatz, R.: A mechanistic model for drug release in PLGA biodegradable stent coatings coupled with polymer degradation and erosion. J. Biomed. Mater. Res., Part A **103**, 2269–2279 (2015)
12. VanEpps, J., Younger, J.: Implantable device related infection. Schock **46**, 597–608 (2016)
13. Wang, M., Tang, T.: Surface treatment strategies to combat implant-related infection from the beginning. J. Orthop. Transl. **17**, 42–54 (2019)

Coupling Temperature with Drug Diffusion: A Second Order Approximation

J. A. Ferreira, Paula de Oliveira, and Elisa Silveira

Abstract The use of enhancers to increase drug release from medical devices and drug transport through tissues has been largely investigated. Researchers from different fields like polymer chemistry, materials science, pharmaceutics, bioengineering, and chemical engineering have addressed efforts to combine materials, stimuli and drugs to design effective drug delivery platforms. For instance heat has been used to increase transdermal drug delivery. Patches with iron batteries are today in the market where heat generated by the batteries increases the drug release from the patches and the permeability of the skin, increasing drug absorption. Heat has been also used to increase drug availability in the target tissue in other contexts like in chemotherapy. In this case, to avoid the side effects of the systemic chemotherapy administration, drugs are encapsulated in thermosensitive carriers that transport the drug to the target where the cargo release is enhanced by heat. The aim of the present work is to study a system of partial differential equations (PDEs), from a numerical point of view, that can been used to describe the drug transport through tissues enhanced by heat. The system is composed by nonlinear PDEs for the temperature and for the drug concentration where the drug diffusion coefficient depends on the temperature. A finite difference method is studied and the qualitative behaviour of the temperature and concentration is numerically illustrated.

Keywords System of partial differential equations · Numerical study · Drug transport through tissues · Finite difference method

1 Introduction

The use of enhancers to increase the drug release from medical devices and the drug transport through the tissues have been largely investigated. Researchers from different fields, like polymer chemistry, materials science, pharmaceutics, bioengineering,

J. A. Ferreira (✉) · P. de Oliveira
CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal
e-mail: ferreira@mat.uc.pt

P. de Oliveira
e-mail: poliveir@mat.uc.pt

E. Silveira
CMUC, University of Coimbra, Coimbra, Portugal
e-mail: elisasilveira11@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_33

and chemical engineering, have been making huge efforts to combine materials, stimuli and drugs properties to design effective drug eluting devices [8, 11, 13, 14, 16]. For instance heat has been used to increase transdermal drug delivery [2, 15] and patches with iron batteries are today in the market. Heat generated by batteries increases the drug release from the patches and the permeability of the skin, leading to a larger drug absorption. Heat has been also used to increase drug availability in chemotherapy [12, 13]. In this case, to avoid the side effects of the systemic chemotherapy administration, drugs are encapsulated in thermoresponsive polymeric transporters that deliver the drugs locally.

Mathematical modeling and numerical simulation is a powerful tool to predict drug release from a medical device and its distribution in the target tissue enhanced by stimuli. The mathematical models combine equations for the stimulus and for the drug transport. To obtain accurate numerical solutions of the qualitative behaviour of the stimulus and drug concentration, it is crucial to use numerical methods with high convergence order, defined on nonuniform grids.

The aim of this work is to study a system of partial differential equations, from a numerical point of view, that can been used to describe drug transport through a tissue enhanced by heat. The system is composed by nonlinear PDEs for the temperature and for the drug concentration, where the drug diffusion coefficient depends on the temperature. Classically the convergence analysis of finite difference methods, for linear initial value problems, is based on the Lax-Richtmyer equivalence theorem that states that a consistent finite difference method is convergent if and only if is stable [9]. From this result, a practical rule used to study convergence of finite difference methods for linear initial value problems, is defined by "Stability and Consistency implies Convergence", where the convergence order at least equal to the consistency order.

For finite difference methods defined on nonuniform meshes, the consistency order may be less than the order of the corresponding finite difference methods defined on uniform meshes. Consequently, based on the Lax rule, it is not possible to conclude to conclude that the convergence order on nonuniform meshes is equal to the convergence order on uniform meshes.

There exists a long list of contributions showing that the convergence order of several linear finite difference methods defined on nonuniform grids is equal to the convergence order of the corresponding finite difference methods defined with uniform grids. Without being exhaustive we mention the classical papers [3, 7, 10], where the analysis requires smoothness of the solutions of the continuous problem, and [1, 4–6] where the convergence analysis requires lower smoothness than those considered in the first group of papers.

Here we are mainly interested in the study of a finite difference scheme for the following system:

$$\frac{\partial T}{\partial t}(x,t) = D_T \frac{\partial^2 T}{\partial x^2}(x,t) + G(T(x,t)),\ (x,t) \in \Omega \times \big(0, T_f\big], \tag{1}$$

and

$$\frac{\partial c}{\partial t}(x,t) = \frac{\partial}{\partial x}\left(D_d(T(x,t))\frac{\partial c}{\partial x}(x,t)\right) + Q(c(x,t)), (x,t) \in \Omega \times \left(0, T_f\right]. \quad (2)$$

The mathematical analysis will be established considering the system (1), (2) completed by the boundary conditions

$$T(t) = 0 \text{ and } c(t) = 0 \text{ on } \partial\Omega \times \left(0, T_f\right] \quad (3)$$

and the initial conditions

$$T(0) = T_0 \text{ and } c(0) = c_0 \text{ in } \Omega \times \left(0, T_f\right]. \quad (4)$$

To simplify the presentation the following notation is used: if $w : \overline{\Omega} \times [0, T_f] \to \mathbb{R}$, by $w(t)$ we represent the function $w(t) : \overline{\Omega} \to \mathbb{R}$ given by $w(t)(x) = w(x,t)$, $x \in \overline{\Omega}$. Here and in the rest of this work $\Omega = (a,b)$, $\partial\Omega = \{a,b\}$, and T_f denotes a final time.

The finite difference method that will be studied is defined on nonuniform grids and it can be seen as a fully discrete piecewise linear finite element method. The convergence analysis will be performed assuming that $T(t), c(t) \in C^4(\overline{\Omega})$, $t \in (0, T_f]$, and we show that the numerical approximations for $T(t)$ and $c(t)$, $T_h(t)$ and $c_h(t)$, respectively, are second order accurate.

This paper is composed by 5 sections. In Sect. 2 we present definitions and basic results. The first convergence results are presented in Sect. 3 where we establish first order estimates for the errors of $T_h(t)$ and $c_h(t)$ assuming that $T(t), c(t) \in C^3(\overline{\Omega})$. These results are improved in Sect. 4 assuming that $T(t), c(t) \in C^4(\overline{\Omega})$, $t \in (0, T_f]$. Finally, in Sect. 5 we present numerical illustrations. To conclude, some conclusions are addressed in Sect. 6.

2 Preliminary Definitions and Results

Let Λ be a sequence of vectors h with positive entries $(h_1, \cdots, h_N)$ such that $\sum_{i=1}^{N} h_i = b - a$ and $h_{max} \to 0$, where $h_{max} = \max_i h_i$. For $h \in \Lambda$, we introduce in $\overline{\Omega}$ the nonuniform grid $\overline{\Omega}_h = \{x_i, i = 0, \cdots, N, x_i - x_{i-1} = h_i, i = 1, \ldots, N, x_0 = a, x_N = b\}$. We denote by Ω_h and $\partial\Omega_h$ the set of interior nodes $\Omega \cap \overline{\Omega}_h$ and the boundary points $\partial\Omega \cap \overline{\Omega}_h$, respectively.

By W_h we represent the space of grid functions defined in $\overline{\Omega}_h$ and the space of grid functions in W_h that are null at the boundary points is denoted by $W_{h,0}$. We introduce in $W_{h,0}$ the inner product $(u_h, v_h)_h = \sum_{i=1}^{N-1} h_{i+1/2} u_h(x_i) v_h(x_i)$, $u_h, v_h \in W_{h,0}$, where $h_{i+1/2} = \frac{h_i + h_{i+1}}{2}$, being the corresponding norm denoted by $\|.\|_h$. We use the notations

$$(u_h, v_h)_+ = \sum_{i=1}^{N} h_i u_h(x_i) v_h(x_i),\ u_h, v_h \in W_h, \qquad \|u_h\|_+ = \Big(\sum_{i=1}^{N} h_i (u_h(x_i))^2\Big)^{1/2}.$$

Let D_{-x}, D_x^* and D_2 be the following finite difference operators:

$$\begin{aligned} D_{-x} u_h(x_i) &= \frac{u_h(x_i) - u_h(x_{i-1})}{h_i},\ i = 1, \cdots, N \\ D_x^* u_h(x_i) &= \frac{u_h(x_{i+1}) - u_h(x_i)}{h_{i+1/2}},\ i = 0, \cdots, N-1, \\ D_2 u_h(x_i) &= \frac{D_{-x} u_h(x_{i+1}) - D_{-x} u_h(x_i)}{h_{i+1/2}}, i = 1, \cdots, N-1, \end{aligned}$$

where $u_h \in W_h$.

In the next result we establish a discrete version of the *integration by parts rule* and a discrete version of Poincaré-Friedrich's inequality.

Proposition 1 *For all $u_h \in W_h$ and $v_h \in W_{h,0}(\overline{\Omega_h})$, we have*

1. $(-D_x^* u_h, v_h)_h = (u_h, D_{-x} v_h)_+$,
2. $(-D_2 u_h, v_h)_h = (D_{-x} u_h, D_{-x} v_h)_+$,
3. $\|u_h\|_h^2 \leq |\Omega| \|D_{-x} u_h\|_+^2$.

where $|\Omega|$ denotes the measure of Ω.

By $T_h(t)$ and $c_h(t)$ we represent the semi-discrete approximations of $T(t)$ and $c(t)$, respectively, defined by the following ordinary differential systems:

$$\begin{cases} T_h'(t) = D_T D_2 T_h(t) + G(T_h(t)) \text{ in } \Omega_h \times \big(0, T_f\big], \\ T_h(t) = 0 \text{ in } \partial\Omega_h \times \big(0, T_f\big], \\ T_h(0) = R_h T_0 \text{ in } \Omega_h, \end{cases} \tag{5}$$

$$\begin{cases} c_h'(t) = D_x^*(D_d(M_h T_h) D_{-x} c_h(t)) + Q(c_h(t)) \text{ in } \Omega_h \times \big(0, T_f\big], \\ c_h(t) = 0 \text{ in } \partial\Omega_h \times \big(0, T_f\big], \\ c_h(0) = R_h c_0 \text{ in } \Omega_h, \end{cases} \tag{6}$$

where $R_h : C(\overline{\Omega}) \to W_h$ is the restriction operator $R_h u(x_i) = u(x_i)$, $i = 0, \ldots, N$, and M_h denotes the average operator $M_h u_h(x_i) = \frac{1}{2}(u_h(x_{i-1}) + u_h(x_i))$, $i = 1, \ldots, N-1$, $u_h \in W_h$.

We notice that $T_h(t)$ and $c_h(t)$ defined by (5) and (6), respectively, can be seen as fully discrete piecewise linear finite element solutions. In fact, the weak formulations of the initial boundary value problems (IBVP) (1), (2), (3), (4) are given by

$$\begin{aligned}&(T'(t),u) = -D_T(\frac{\partial T}{\partial x}(t),u') + (G(T(t)),u) \text{ a.e. in } (0,T_f], \forall u\in H_0^1(\Omega),\\ &(T(0),u) = (T_0,u),\ \forall u \in L^2(\Omega),\end{aligned} \tag{7}$$

and

$$\begin{aligned}&(c'(t),w) = -(D_d(T(t))\frac{\partial c}{\partial x}(t),w') + (Q(c(t)),w) \text{ a.e. in } (0,T_f], \forall w\in H_0^1(\Omega),\\ &(c(0),w) = (c_0,w),\ \forall w \in L^2(\Omega).\end{aligned} \tag{8}$$

In (7), (8), *a.e.* means *almost everywhere* and by $L^2(\Omega)$ and $H_0^1(\Omega)$ we denote the usual Sobolev spaces endowed with the usual inner products and norms.

The piecewise linear finite element approximations for $T(t)$ and $c(t)$ defined by (7) and (8), respectively, are computed considering the piecewise linear interpolation functions $P_hT_h(t)$, $P_hc_h(t) \in H_0^1(\Omega)$ of $T_h(t)$, $c_h(t) \in W_{h,0}$, respectively. These functions are solutions of the following weak problems:

$$\begin{aligned}&(P_hT_h'(t),P_hu_h) = -D_T(\frac{\partial P_hT_h}{\partial x}(t),P_hu_h') + (G(P_hT_h(t)),P_hu_h)\text{in } (0,T_f], \forall u_h\in W_{h,0},\\ &(P_hT_h(0),P_hu_h) = (P_hR_hT_0,P_hu_h),\ \forall u_h \in W_{h,0},\end{aligned} \tag{9}$$

and

$$\begin{aligned}&(P_hc_h'(t),P_hw_h) = -(D_d(P_hT_h(t))\frac{\partial P_hc_h}{\partial x}(t),P_hw_h') + (Q(P_hc_h(t)),P_hw_h) \text{ in } (0,T_f], \forall w_h\in W_{h,0},\\ &(P_hc_h(0),P_hw_h) = (P_hR_hc_0,P_hw_h),\ \forall w_h \in W_{h,0}.\end{aligned} \tag{10}$$

The two finite problems (9), (10) are then replaced by the fully discrete piecewise linear finite element approximations

$$\begin{aligned}&(T_h'(t),u_h)_h = -D_T(D_{-x}T_h(t),D_{-x}u_h)_+(G(T_h(t)),u_h)_h\text{in } (0,T_f], \forall u_h\in W_{h,0},\\ &(T_h(0),u_h)_h = (R_hT_0,u_h)_h,\ \forall u_h \in W_{h,0},\end{aligned} \tag{11}$$

and

$$\begin{aligned}&(c_h'(t),w_h)_h = -(D_d(M_hT_h(t))D_{-x}c_h(t),D_{-x}w_h)_+ + (Q(c_h(t)),w_h)_h \text{ in } (0,T_f], \forall w_h\in W_{h,0},\\ &(c_h(0),w_h)_h = (R_hc_0,w_h)_h,\ \forall w_h \in W_{h,0}.\end{aligned} \tag{12}$$

Finally, choosing in each equation of (11), (12) a sequence of grid functions where each element is equal to one in a grid point and zero in the rest we arrive to the IBVP (5) and (6).

3 Convergence Analysis for Solutions in $C^3(\overline{\Omega})$

In this section we establish estimates for the errors $E_T(t) = R_hT(t) - T_h(t)$, $E_c(t) = R_hc(t) - c_h(t)$, where $T_h(t), c_h(t)$ are given by (5) and (6) or (11) and (12). We assume that the coefficient function D_d satisfies the following assumption:
$H_1 : D_d \in C^1_b(\mathbb{R})$ and $D_d \geq \beta \geq 0$ in $\mathbb{R}$.
We assume that G and Q have bounded first order derivatives, that is

$$\max_{\mathbb{R}} G' \leq C_G, \quad \max_{\mathbb{R}} Q' \leq C_Q,$$

where C_G and C_Q are constants.

3.1 Temperature

We start by studying the error $E_T(t) = R_hT(t) - T_h(t)$.

Theorem 1 *Let the solution T of (1) in $L^2(0, T_f, C^3(\overline{\Omega}))$ and let T_h be defined by (5), such that $R_hT, T_h \in C^1([0, T_f], W_{h,0})$. Then there exists a positive constant $Const$, h and t independent, such that $E_T(t) = R_hT(t) - T_h(t)$ satisfies*

$$\|E_T(t)\|_h^2 + \int_0^t e^{C_G(t-s)}\|D_{-x}E_T(s)\|_+^2 ds \leq Const h^2_{max} \int_0^t e^{C_G(t-s)}\|T(s)\|^2_{C^3(\overline{\Omega})} ds, \tag{13}$$

for $t \in [0, T_f]$ and $h \in \Lambda$.

Proof Let $T_{r,T}(t)$ be the truncation error induced by the spatial discretization defined in (5). For $T_{r,T}(t)$ we have the following representation

$$T_{r,T}(x_i, t) = \frac{1}{6}\Big(h^2_{i+1}\frac{\partial^3 T}{\partial x^3}(\eta_i, t) - h_i^2\frac{\partial^3 T}{\partial x^3}(\xi_i, t)\Big),$$

where $\eta_i, \xi_i \in [x_{i-1}, x_{i+1}]$, $i = 1, \ldots, N-1$.

It can be shown that for the error $E_T(t)$ we have

$$\begin{aligned}(E_T(t)', E_T(t))_h &= -D_T(D_{-x}E_T(t), D_{-x}E_T(t))_+ + (G(R_hT(t)) - G(T_h(t)), E_T(t))_h \\ &\quad + (T_{r,T}(t), E_T(t))_h,\ t \in (0, T_f].\end{aligned}$$

Young's inequality leads to

$$\frac{1}{2}\frac{d}{dt}\|E_T(t)\|_h^2 + D_T\|D_{-x}E_T(t)\|_+^2 \leq \frac{1}{4\varepsilon^2} \leq \|T_{r,T}(t)\|_h^2 + \varepsilon^2\|E_T(t)\|_h^2 + C_G\|E_T(t)\|_h^2,$$

where $\varepsilon \neq 0$. Considering now the discrete Poincaré-Friedrich's inequality we get

$$\frac{d}{dt}\|E_T(t)\|_h^2 + 2(D_T - |\Omega|\varepsilon^2)\|D_{-x}E_T(t)\|_+^2 + \frac{1}{2\varepsilon^2}\|T_{r,T}(t)\|_h^2 + C_G\|E_T(t)\|_h^2. \tag{14}$$

To establish an error estimation for $E_T(t)$, we need to compute an upper bound for $\|T_{r,T}(t)\|_h^2$. As

$$\|T_{r,T}(t)\|_h^2 \le \frac{2}{9}D_T^2|\Omega|\|T(t)\|_{C^3(\overline{\Omega})}h_{max}^2,$$

fixing ε such that $D_T - \varepsilon^2|\Omega| > 0$, and defining $Const = \dfrac{1}{18\varepsilon^2}\dfrac{D_T^2|\Omega|}{D_T - \varepsilon^2|\Omega|}$, we conclude that

$$\frac{d}{dt}\Big(e^{-C_G t}\|E_T(t)\|_h^2 + \int_0^t e^{-C_G s}\|D_{-x}E_T(s)\|_+^2 ds - Const\int_0^t e^{-C_G s}\|T(s)\|_{C^3(\overline{\Omega})ds}\Big) \le 0, \tag{15}$$

for $t \in [0, T_f]$. This inequality leads to (13).

3.2 Concentration

In this section we establish an upper bound for the error $E_c(t) = R_h c(t) - c_h(t)$, where c is defined by (2). As c depends on the solution T of (1), we will get an upper bound for $E_c(t)$ depending on the error $E_T(t)$ as well as on the truncation error associated with the spatial discretization defined in (6).

Theorem 2 *Let T and c be solutions of (1) and (2), respectively, in $L^2(0, T_f, C^3(\overline{\Omega}))$ and let T_h, c_h be defined by (5) and (6), respectively. Let $E_T(t)$ and $E_c(t)$ be the spatial-discretization errors $E_T(t) = R_h T(t) - T_h(t)$ and $E_c(t) = R_h c(t) - c_h(t)$. If $R_h c, c_h \in C^1([0, T_f], W_{h,0})$ and $R_h T, T_h \in C([0, T_f], W_{h,0})$, there exists a positive constant $Const$, h and t independent, such that*

$$\begin{aligned}\|E_c(t)\|_h^2 + \int_0^t e^{C_Q(t-s)}\|D_{-x}E_c(s)\|_+^2 ds \le Const\Big(&\int_0^t e^{C_Q(t-s)}\|E_T(s)\|_h^2\|c(s)\|_{C^1(\overline{\Omega})}^2 ds \\ &+ h_{max}^2\int_0^t e^{(C_Q+C_G)(t-s)}\|c(s)\|_{C^3(\overline{\Omega})}^2(\|T(s)\|_{C^2(\overline{\Omega})}^2 + 1)ds + h_{max}^2\Big),\end{aligned} \tag{16}$$

for $t \in [0, T_f]$ and $h \in \Lambda$.

Proof Let $T_{r,c}(t)$ be the truncation error induced by the spatial discretization defined in (6). It can be shown that $T_{r,c}(t)$ admits the representation

$$\begin{aligned}T_{r,c}(x_i,t) = {} & D_d'(T(x_i,t))\left(\frac{h_i - h_{i+1}}{2}\right)\left[\frac{\partial^2 T}{\partial x^2}(x_i,t)\frac{\partial c}{\partial x}(x_i,t) + \frac{\partial T}{\partial x}(x_i,t)\frac{\partial^2 c}{\partial x^2}(x_i,t)\right] \\ & + \frac{D_d(T(x_i,t))}{6h_{i+1/2}}\left(h_i^2\frac{\partial^3 c}{\partial x^3}(\xi_i,t) - h_{i+1}^2\frac{\partial^3 c}{\partial x^3}(\eta_i,t)\right) + \mathcal{O}(h_{max}^2),\end{aligned} \tag{17}$$

where $\mathscr{O}(h_{max}^2)$ represents a term, depending on $\|c(t)\|_{C^3(\overline{\Omega})}$ and $\|T(t)\|_{C^3(\overline{\Omega})}$, such that $|\mathscr{O}(h_{max}^2)| \leq Const\, h_{max}^2$, being $Const$ a positive and h and time independent constant.

From (17), for $\|Tr_c(t)\|_h^2$ we easily get the following

$$\|T_{r,c}(t)\|_h^2 \leq Consth_{max}^2\Big((\|T(t)\|^2_{C^2(\overline{\Omega})} + 1)\|c(t)\|^2_{C^3(\overline{\Omega})} + h_{max}^2\Big). \tag{18}$$

For the error $E_c(t)$ we have

$$\begin{aligned}(E_c(t)', E_c(t))_h = &-([D_d(M_hR_hT(t)) - D_d(M_hT_h(t))]\, D_{-x}R_hc(t), D_{-x}E_c(t))_+ \\ &- ((D_d(M_hT_h(t)))D_{-x}E_c(t), D_{-x}E_c(t))_+ \\ &+ (Q(R_hc(t)) - Q(c_h(t)), E_c(t))_h + (T_{r,c}(t), E_c(t))_h.\end{aligned} \tag{19}$$

As

$$\begin{aligned}&|([D_d(M_hR_hT(t)) - D_d(M_hT_h(t))]\, D_{-x}R_hc(t), D_{-x}E_c(t))_+| \\ &\qquad \leq \|D_d\|_{C_b^1(\mathbb{R})}\sqrt{2}\|E_T(t)\|_h\|c(t)\|_{C^1(\overline{\Omega})}\|D_{-x}E_c(t)\|_+,\end{aligned}$$

for $\varepsilon_i \neq 0, i = 1, 2$, considering the assumption H_1 we obtain

$$\begin{aligned}&\frac{d}{dt}\|E_c(t)\|_h^2 + 2(\beta - \varepsilon_1^2 - |\Omega|\varepsilon_2^2)\|D_{-x}E_c(t)\|_+^2 \\ &\leq \frac{1}{\varepsilon_1^2}\|D_d\|_{C_b^1(\mathbb{R})}\|E_T(t)\|_h^2\|c(t)\|^2_{C^1(\overline{\Omega})} + \frac{1}{2\varepsilon_2^2}\|T_{r,c}(t)\|_h^2 + C_Q\|E_c(t)\|_h^2.\end{aligned} \tag{20}$$

Then, fixing the constants $\varepsilon_i, i = 1, 2$, such that $2(\beta - \varepsilon_1^2 - |\Omega|\varepsilon_2^2) > 0$, we guarantee the existence of a positive constant $Const$, h and t independent, such that

$$\begin{aligned}&\|E_c(t)\|_h^2 + \int_0^t e^{C_Q(t-s)}\|D_{-x}E_c(s)\|_+^2 ds \\ &\leq Const \int_0^t e^{C_Q(t-s)}\Big(\|E_T(s)\|_h^2\|c(s)\|^2_{C^1(\overline{\Omega})} + \|T_{r,c}(s)\|_h^2\Big)ds,\end{aligned} \tag{21}$$

and taking into account the upper bound (18) we deduce (16).

4 Convergence Analysis for Solutions in $C^4(\overline{\Omega})$

In Theorems 1 and 2, assuming that the solutions $T(t)$ and $c(t)$ are in $C^3(\overline{\Omega})$, we establish that

$$\|E_T(t)\|_h \leq Consth_{max}, \|E_c(t)\|_h \leq Consth_{max},$$

and

$$\int_0^t e^{C_G(t-s)}\|D_{-x}E_T(s)\|_+^2 ds \leq Const h_{max}^2, \int_0^t e^{C_Q(t-s)}\|D_{-x}E_c(s)\|_+^2 ds \leq Const h_{max}^2.$$

In this section we increase the convergence orders increasing the regularity of $T(t)$ and $c(t)$, namely by assuming that $T(t), c(t) \in C^4(\overline{\Omega})$.

4.1 Temperature

Theorem 3 *Let T be solution of (1) in $L^2(0, T_f, C^4(\overline{\Omega}))$ and let T_h be defined by (5), such that $R_hT, T_h \in C^1([0, T_f], W_{h,0})$. Then there exists a positive constant $Const$, h and t independent, such that $E_T(t) = R_hT(t) - T_h(t)$ satisfies*

$$\|E_T(t)\|_h^2 + \int_0^t e^{C_G(t-s)}\|D_{-x}E_T(s)\|_+^2 ds \leq Const h_{max}^4 \int_0^t e^{C_G(t-s)}\|T(s)\|_{C^4(\overline{\Omega})}^2 ds, \tag{22}$$

for $t \in [0, T_f]$ and $h \in \Lambda$.

Proof As in the proof of Theorem 1, we have

$$\frac{1}{2}\frac{d}{dt}\|E_T(t)\|_h^2 + D_T\|D_{-x}E_T(t)\|_+^2 = (T_{r,T}(t), E_T(t))_h + (G(R_hT(t)) - G(T_h(t)), E_T(t))_h. \tag{23}$$

Taking into account that $T(t) \in C^4(\overline{\Omega})$, we have for $T_{r,T}(t)$ the following representation

$$T_{r,T}(x_i, t) = \frac{D_T}{3}(h_{i+1} - h_i)\frac{\partial^3 T}{\partial x^3}(x_i, t) + O(h_{max}^2),$$

where $|O(h_{max}^2)| \leq Const h_{max}^2\|T(t)\|_{C^4(\overline{\Omega})}$. Then for $(T_{r,T}(t), E_T(t))_h$ we obtain

$$(T_{r,T}(t), E_T(t))_h = \frac{D_T}{6}\sum_{i=1}^{N-1}\left(\frac{\partial^3 T}{\partial x^3}(x_i, t)E_T(x_i, t) - \frac{\partial^3 T}{\partial x^3}(x_{i-1}, t)E_T(x_{i-1}, t)\right)$$
$$+ (O(h_{max}^2), E_T(t))_h,$$

that leads to

$$\begin{aligned}(T_{r,T}(t), E_T(t))_h &\leq \frac{D_T}{6}\sum_{i=1}^{N-1} h_i^3\frac{\partial^3 T}{\partial x^3}(x_i, t)D_{-x}E_T(x_i) + \frac{D_T}{6}\sum_{i=1}^{N-1} h_i^2\int_{x_{i-1}}^{x_i}\frac{\partial^4 T}{\partial x^4}(x, t)dx E_T(x_{i-1}, t)\\ &+ (O(h_{max}^2), E_T(t))_h\\ &= \frac{D_T}{6}(A + B) + (O(h_{max}^2), E_T(t))_h.\end{aligned}$$

For A and B we have the following upper bounds

$$|A| \leq h_{max}^2\sqrt{|\Omega|}\|T(t)\|_{C^3(\overline{\Omega})}\|D_{-x}E_T(t)\|_+ \tag{24}$$

and

$$|B| \leq h_{max}^2 \sqrt{2} \| \frac{\partial^4 T}{\partial x^4}(t)\|_{L^2(\Omega)} \|E_T(t)\|_h, \tag{25}$$

respectively.

Then, considering Young's and Poincaré-Friedrich's inequalities and $\varepsilon_i \neq 0, i = 1, 2, 3$, we get

$$\begin{aligned} |(T_{r,T}(t), E_T(t))_h| \leq h_{max}^4 \Big(\frac{D_T^2}{36} \Big(\frac{1}{4\varepsilon_1^2} |\Omega| \|T(t)\|^2_{C^3(\overline{\Omega})} + \frac{1}{\varepsilon_2^2} \| \frac{\partial^4 T}{\partial x^4}(t)\|^2_{L^2(\Omega)} \Big) \\ + Const \frac{1}{\varepsilon_3^2} \|T(t)\|^2_{C^4(\overline{\Omega})} \Big) + (\varepsilon_1^2 + |\Omega|(\varepsilon_2 + \varepsilon_3^2)) \|D_{-x} E_T(t)\|_+^2. \end{aligned} \tag{26}$$

Taking the last upper bound in (23) we deduce

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}\|E_T(t)\|_h^2 + (D_T - (\varepsilon_1^2 + |\Omega|(\varepsilon_2 + \varepsilon_3^2)))\|D_{-x}E_T(t)\|_+^2 \\ &\leq h_{max}^4 \Big(\frac{D_T^2}{36} \Big(\frac{1}{4\varepsilon_1^2} |\Omega| \|T(t)\|^2_{C^3(\overline{\Omega})} + \frac{1}{\varepsilon_2^2} \| \frac{\partial^4 T}{\partial x^4}(t)\|^2_{L^2(\Omega)} \Big) + Const \frac{1}{\varepsilon_3^2} \|T(t)\|^2_{C^4(\overline{\Omega})} \Big) \\ &+ C_G \|E_T(t)\|_h^2, \end{aligned} \tag{27}$$

for $t \in (0, T_f]$. Fixing in (27) $\varepsilon_i \neq 0, i = 1, 2, 3$, such that $D_T - (\varepsilon_1^2 + |\Omega|(\varepsilon_2 + \varepsilon_3^2)) > 0$, we guarantee the existence a positive constant $Const$, h and t independent, such that (22) holds.

4.2 Concentration

Theorem 4 *Let T and c be solutions of (1) and (2), respectively, in $L^2(0, T_f, C^4(\overline{\Omega}))$ and let T_h, c_h be defined by (5) and (6), respectively. Let $E_T(t)$ and $E_c(t)$ be the spatial-discretization errors $E_T(t) = R_h T(t) - T_h(t)$ and $E_c(t) = R_h c(t) - c_h(t)$. If $R_h c, c_h \in C^1([0, T_f], W_{h,0})$ and $R_h T, T_h \in C([0, T_f], W_{h,0})$, there exists a positive constant h and t independent such that*

$$\begin{aligned} \|E_c(t)\|_h^2 + \int_0^t e^{C_Q(t-s)} \|D_{-x}E_c(s)\|_+^2 ds \leq Const \Big(\int_0^t e^{C_Q(t-s)} \|E_T(s)\|_h^2 \|c(s)\|^2_{C^1(\overline{\Omega})} ds \\ + h_{max}^4 \int_0^t e^{C_Q(t-s)} \|c(s)\|^2_{C^4(\overline{\Omega})} (\|T(s)\|^2_{C^3(\overline{\Omega})} + 1) ds \Big), \end{aligned} \tag{28}$$

for $t \in [0, T_f]$ and $h \in \Lambda$.

Proof The truncation error induced by the spatial discretization defined in (6) has the representation

$$\begin{aligned} T_{r,c}(x_i,t) &= (h_i - h_{i+1})\Big(\frac{1}{3}D'_d(T(x_i,t))\left[\frac{\partial^2 T}{\partial x^2}(x_i,t)\frac{\partial c}{\partial x}(x_i,t) + \frac{\partial T}{\partial x}(x_i,t)\frac{\partial^2 c}{\partial x^2}(x_i,t)\right] \\ &\quad + \frac{1}{3}D_d(T(x_i,t))\frac{\partial^3 c}{\partial x^3}(x_i,t)\Big) + \mathscr{O}(h_{max}^2), \end{aligned}$$

where $|\mathscr{O}(h_{max}^2)| \leq Const\|c(t)\|_{C^4(\overline{\Omega})}\big(\|T(t)\|_{C^3(\overline{\Omega})} + 1\big)$.

Let $g(x,t)$ be defined by

$$\begin{aligned} g(x,t) = &\ \Big(\frac{1}{3}D'_d(T(x,t))\left[\frac{\partial^2 T}{\partial x^2}(x,t)\frac{\partial c}{\partial x}(x,t) + \frac{\partial T}{\partial x}(x,t)\frac{\partial^2 c}{\partial x^2}(x,t)\right] \\ &+ \frac{1}{3}D_d(T(x,t))\frac{\partial^3 c}{\partial x^3}(x,t)\Big). \end{aligned}$$

Then, as in the proof of the Theorem 3, we have

$$\begin{aligned} |(T_{r,c}(t), E_c(t))_h| &\leq \frac{1}{2}h_{max}^2\Big(\|g(t)\|_{C(\overline{\Omega})}\sqrt{|\Omega|}\|D_{-x}E_c(t)\|_+ + \sqrt{2}\|\frac{\partial g}{\partial x}(t)\|_{L^2(\Omega)}\|E_c(t)\|_h\Big) \\ &\quad + \|\mathscr{O}(h_{max}^2)\|\|E_c(t)\|_h. \end{aligned}$$

Consequently, discrete Poincaré-Friedrichs inequality leads to

$$\begin{aligned} |(T_{r,c}(t), E_c(t)| \leq &\ \frac{1}{2\varepsilon_1^2}|\Omega|\Big(h_{max}^4\Big(\|g(t)\|^2_{C(\overline{\Omega})} + 2\|\frac{\partial g}{\partial x}(t)\|^2_{L^2(\Omega)}\Big) + \|\mathscr{O}(h_{max}^2)\|_h^2\Big)^2 \\ &+ \varepsilon_1^2\|D_{-x}E_c(t)\|_+^2, \end{aligned}$$

where $\varepsilon_1 \neq 0$.

Following the proof of the Theorem 2, we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|E_T(t)\|_h^2 + (\beta - \varepsilon_1^2 - \varepsilon_2^2)\|D_{-x}E_T(t)\|_+^2 \leq &\ \frac{1}{2\varepsilon_2^2}\|D_d\|_{C_b^1(\mathbb{R})}\|E_T(t)\|_h^2\|c(t)\|^2_{C^1(\overline{\Omega})} \\ &+ \frac{1}{2\varepsilon_1^2}|\Omega|\Big(h_{max}^4\Big(\|g(t)\|^2_{C(\overline{\Omega})} + 2\|\frac{\partial g}{\partial x}(t)\|^2_{L^2(\Omega)}\Big) + \|\mathscr{O}(h_{max}^2)\|_h^2\Big)^2 \\ &+ C_Q\|E_c(t)\|_h^2. \end{aligned} \tag{29}$$

Fixing $\varepsilon_i \neq 0, i = 1, 2$, such that $\beta - \varepsilon_1^2 - \varepsilon_2^2 > 0$, we guarantee the existence of a positive constant $Const$, h and t independent, such that

$$\begin{aligned} \frac{d}{dt}\|E_c(t)\|_h^2 + \|D_{-x}E_c(t)\|_+^2 \leq &\ Const\Big(\|E_T(t)\|_h^2\|c(t)\|^2_{C^1(\overline{\Omega})} \\ &+ h_{max}^4\|c(t)\|^2_{C^4(\overline{\Omega})}(\|T(t)\|^2_{C^3(\overline{\Omega})} + 1)\Big) + C_Q\|E_c(t)\|_h^2, t \in (0, T_f]. \end{aligned} \tag{30}$$

Finally, to conclude, we remark that inequality (30) leads to (28).

From Theorems 3 and 4 we conclude the next result:

Corollary 1 *Under the assumptions of the Theorems 3 and 4 for the error $E_c(t) = R_hc(t) - c_h(t)$ holds the following*

$$\begin{aligned}&\|E_c(t)\|_h^2 + \int_0^t e^{C_Q(t-s)}\|D_{-x}E_c(s)\|_+^2 ds\\ &\le Consth_{max}^4\int_0^t \Big(\|T(s)\|^2_{C^4(\overline{\Omega})} + \|c(s)\|^2_{C^4(\overline{\Omega})}(\|T(s)\|^2_{C^3(\overline{\Omega})}+1)\Big)ds.\end{aligned}$$

for $t \in [0, T_f]$ and $h \in \Lambda$.

5 Numerical Simulation

In this section we illustrate the qualitative behaviour of T and c combining (5) and (6) with an explicit-implicit approach defined using Euler's method. The concentration diffusion term is approximated implicitly and the reaction terms in the temperature and concentration equations are approximately explicitly. We take $\overline{\Omega} = [0, 1]$, $h_{max} = 10^{-2}$, $[0, T_f] = [0, 10]$, the time stepsize $\Delta t = 10^{-2}$, $D_d(T(t)) = D_0e^{-\frac{K}{T(t)}}$, $D_0 = 10^{-2}$, $K = 10/8.314$, $D_T(T) = D_d(T), v(T(t)) = bT(t)$, with $b = 10^{-1}$, $T(0, x) = 310x(1-x)(K)$, $c(c, 0) = x(1-x)$, $x \in [0, 1]$, G defined by

$$G(T(t)) = 312 - 2cos(3t), \text{ for} \in (0, 0.4), \text{ and } G(T(t)) = 0 \text{ for } t \ge 0.4, \quad (31)$$

and $Q = 0$.

Figure 1 illustrates the behaviour of T (left) and c(right) at $t = 1$. These plots were obtained with and without the heat source G defined by (31). We observe that at $t = 1$, the drug concentration in Ω under the effect of the heat source is approximately zero. However a significant amount of drug remains in Ω when no heat enhancement is considered. These results show the efficacy of the stimulus in the drug delivery. To confirm our conclusions, in Fig. 2 we illustrate the behaviour of the mass released $M_r(t) = M(0) - \displaystyle\int_\omega c(x, t)dx$, $t \in [0, 4]$, where $M(0) = \int_\Omega c(x, 0)dx$ is the initial drug mass. As it can be seen, with the source term G, M_r attains its stationary state near to $t = 1$ while, without the heat source, $M_r(t)$ attains the stationary state near to $t = 3.5$.

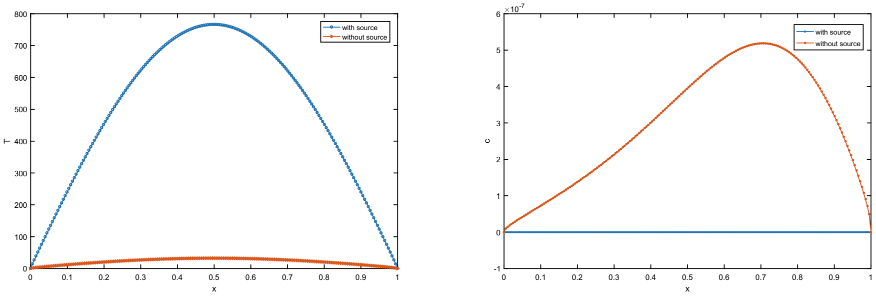
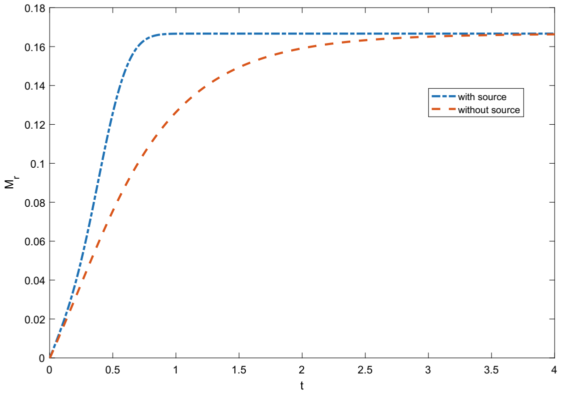


Fig. 1 The plots of T (left) and c (right) at $t = 1$ with and without the heat source G defined by (31)

Fig. 2 The plots of the released mass $M_r(t)$ with and without the heat source G defined by (31)



6 Conclusion

In this work we establish error estimates for the numerical approximations of (1) and (2) defined by the finite difference methods (5) and (6) on nonuniform grids. These methods can be seen as fully discrete piecewise linear finite element methods.

Theorems 3 and 4 are the main results of this work. In Theorem 3 we prove that the finite difference method (5) leads to second order approximations for the solution of (1). This result shows that the method is supraconvergent: though the spatial truncation error is only of first order, the method is second order convergent. In Theorem 4 we establish an error estimate for c_h defined by (6) when T_h is given by (5). This result allows us to conclude that the finite difference method (5), (6) is also second-order convergent. The regularity of the solutions of the IBVP (1) and (2) is the main requirement imposed in the proof of the mentioned results, $T(t), c(t) \in C^4(\overline{\Omega})$.

The qualitative behaviour of the solutions of the IBVPs (1), (2), (3) and (4) is numerically illustrated. Let t_T^* denotes the time needed for the drug released mass $M_r(t)$ to reach its stationary state when heat is used to enhance the drug transport. Let t^* be the corresponding time without the action of the stimulus. The numerical results presented in Fig. 2 show that $t^*_T << t^*$. This finding confirms that heat can be an effective stimulus to enhance drug transport.

Acknowledgements This work was supported by Centro de Matemática da Universidade de Coimbra UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MCTES and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020. The first and the second authors were also supported by the project *NEXT.parts - Next- generation of advanced hybrid parts*, funded by EU's Horizon 2020 science programme (Portugal 2020, COMPETE 2020).

Elisa Silveira was supported by the FCT PhD grant PD/BD/128058/2016 funded by the Portuguese Government through FCT/MCTES.

References

1. Barbeiro, S., Ferreira, J.A., Grigorieff, R.D.: Supraconvergence of a finite difference scheme for solutions in $H^s(0, l)$. IMA J. Numer. Anal. **25**, 797–811 (2005)
2. Evanghelidis, A., Beregoi, M., Diculescu, V., Galatanu, A., Ganea, P., Enculescu, I.: Delivery patch systems based on thermoresponsive hydrogels and submicronic fiber heaters. Sci. Rep. **8**, 17555 (2018)
3. Ferreira, J.A., Grigorieff, R.: On the supraconvergence of elliptic finite difference schemes. Appl. Numer. Math. **28**, 275–292 (1998)
4. Ferreira, J.A., Grigorieff, R.: Supraconvergence and supercloseness of a scheme for elliptic equations on non-uniform grids. Numer. Funct. Anal. Optim. **27**(5–6), 539–564 (2006)
5. Ferreira, J.A., de OLiveira, P., Silveira, E.: Drug release enhanced by temperature: an accurate discrete model for solutions in H^3. Comput. Math. Appl. (to appear). https://doi.org/10.1016/j.camwa.2019.08.002
6. Ferreira, J.A., Pinto, L.: Supraconvergence and supercloseness in quasilinear coupled problems. J. Comput. Appl. Math. **252**, 120–131 (2013)
7. Forsyth, P., Sammon, P.H.: Quadratic convergence for cell-centered grids. Appl. Numer. Math. **4**, 377–394 (1988)
8. Mura, S., Couvreur, P.: Stimuli-responsive nanocarriers for drug delivery. Nat. Mater. **12**, 991–1003 (2013)
9. Lax, P., Richtmyer, R.: Survey of the stability of linear finite difference equations. Commun. Pure Appl. Math. **9**, 267–293 (1956)
10. Manteuffel, T., White Jr., A.: The numerical solution of second order boundary value problems on nonuniform meshes. Math. Comput. **47**, 511–535 (1986)
11. Patra, J., Das, G., Fraceto, L., Campos, E., Rodriguez-Torres, M., Acosta-Torres, L., Diaz-Torres, L., Grillo, R., Swamy, M., Sharma, S., Habtemariam, S., Shin, H.: Nano based drug delivery systems: recent developments and future prospects. J. Nanobiotechnol. **16**, 71 (2018)
12. Qiao, Y., Wan, J., Zhou, L., Ma, W., Yang, Y., Luo, W., Yu, Z., Wang, H.: Stimuli-responsive nanotherapeutics for precision drug delivery and cancer therapy. WIREs Nanomed. Nanobiotechnol. **11**, e1527 (2019)
13. Senapati, S., Mahanta, A., Kumar, S., Maiti, P.: Controlled drug delivery vehicles for cancer treatment and their performance. Signal Transd. Targeted Ther. **3**, 7 (2018)
14. Sahle, F., Gulfam, M., Lowe, T.: Design strategies for physical stimuli-responsive programmable nanotherapeutics. Drug Discov. Today **23**, 992–1006 (2018)
15. Szunerits, S., Boukherroub, R.: Heat: a highly efficient skin enhancer for transdermal drug delivery. Fontier Bioeng. Biotechnol. **6**, 15 (2018)
16. Wells, C., Harris, M., Choi, L., Murali, V., Guerra, F., Jennings, J.: Stimuli-responsive drug release from smart polymers. J. Funct. Biomater. **10**, 34 (2019)

Drug Release from Thermosensitive Polymeric Platforms—Towards Non Fickian Models

J. A. Ferreira, Paula de Oliveira, and Elisa Silveira

Abstract To reduce the side effects of chemotherapy drugs, namely in cancer therapy, researchers of different fields are making tremendous efforts to design new drug systems that can be used to deliver the drug locally in a sustainable way. Polymeric nanocarriers are investigated to transport the drug to the target tissue where the cargo should be delivered. Physical or chemical stimuli are being considered to enhance the drug release. Heat is investigated to enhance drug release and drug transport in different scenarios as transdermal drug delivery or cancer treatment. In this case the drug is entrapped in a thermoresponsive polymeric carrier and its release is enhanced by the temperature increasing. Traditionally, drug transport and the temperature evolution are described by parabolic equations based on Fick's law. To describe accurately drug transport through a viscoelastic material, Fick's law should be modified to include the viscoelastic properties of the tissues. In this work we propose a mathematical model to describe the drug release from a thermoresponsive platform. To simplify, we assume that the time evolution occurs at discrete time levels that allows the introduction of the temperature as an input of our model. An explicit representation of the drug concentration is obtained.

Keywords Fickian models · Chemotherapy drugs · Thermosensitive polymeric platforms · Partial differential equations

J. A. Ferreira (✉) · P. de Oliveira
Department of Mathematics, University of Coimbra, CMUC, Coimbra, Portugal
e-mail: ferreira@mat.uc.pt

P. de Oliveira
e-mail: poliveir@mat.uc.pt

E. Silveira
University of Coimbra, CMUC, Coimbra, Portugal
e-mail: elisasilveira11@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_34

1 Introduction

To reduce the side effects of chemotherapy drugs, namely in cancer therapy, pharmaceutics, material engineers, medical researchers, are addressing their effort to design new drug delivery systems that can be used to release the drug locally. When drug delivery platforms are in contact with the target tissue, the drug release can be stimulated by physical or chemical stimuli like heat, light, ultrasound, electric or magnetic fields, pH or enzymes. We mention without being exhaustive the recent papers [1, 9, 10] and the references therein. To put the drug in contact with the target, the drug is entrapped in polymeric platforms and the release occurs under the action of a stimulus. It is clear that the polymeric structure should react in the presence of the stimulus.

Different biodegradable and biocompatible polymers are currently used in drug release systems. PNIPAAm, Poly(N-isopropylacrylamide), is a thermoresponsive polymer extensively investigated for applications on controlled delivery [8] and in cancer therapy [2]. The drug release from a thermoresponsive polymeric platform, or from a stimuli responsive polymeric structure, is the corollary of a cascade of phenomena: fluid entrance, polymer swelling, drug dissolution and drug transport.

The classical diffusion equations

$$\frac{\partial T}{\partial t} = D_T \Delta T + G(T) \text{ in } \Omega \times \left(0, T_f\right] \tag{1}$$

and

$$\frac{\partial c}{\partial t} = \nabla \left(D_d(T) \nabla c\right) + Q(c) \text{ in } \Omega \times \left(0, T_f\right], \tag{2}$$

for the temperature T and drug concentration c, are often used to describe the transport of a drug enhanced by heat in a domain Ω representing a polymeric reservoir, a target tissue or both domains coupled. In (1), and (2), D_T and D_d stand for the diffusion coefficients of the temperature and the drug respectively. These equations are established using Fick's law for the flux J

$$J(x, t) = -D \nabla \ell(x, t) \tag{3}$$

and the mass conservation equation

$$\frac{\partial \ell}{\partial t} + \nabla J = R(\ell) \text{ in } \Omega \times (0, T_f], \tag{4}$$

where $\ell = T, c$ and $R = G, Q$ represents the reaction terms.

We notice that both media, polymeric platform and target tissue, can be seen as viscoelastic materials where the drug transport or the fluid entrance is of non-Fickian type. In fact, when the fluid or drug molecules move through the medium, a resistance to particles movement is imposed by the medium structure. This resistance can be seen as a stress driven diffusion that acts as a barrier to particles transport. In [4, 5] and [6], the resistance offered by the internal medium structure is interpreted as an opposite convective field. We remark that hyperbolic equations of wave type were proposed to describe the heat transport in [3] and studied in [7].

Stimuli responsive polymers, like PNIPAAm, are characterized by a phase transition temperature, the so called *Critical Solution Temperature* (CST): they swell for temperatures below the CST and shrink when the temperature is above the CST [12]. The change of state of the polymer is a continuous function of the temperature. The mathematical modeling of this continuous change requires a mathematical law for the polymer shape evolution in function of the temperature. To the best of our knowledge the issue has not been described in the literature.

The mathematical modeling of drug release from a temperature sensitive polymer was considered in [11] using a Fickian description for the drug transport. The author assumes that the temperature effects on the polymer states occur discretely in time. In this last work, the spatial domain has two different configurations and the drug transport is described by the diffusion Eq. (2). In the present paper, while studying the effect of the temperature on the behaviour of the polymer as in [11], the viscoelastic effect of the polymeric structure on the drug transport is taken into account. We assume that the polymer has two different states: the swollen and the shrink states that change discretely in time. Each one of these states is characterized by different Young modulus, defined by the intensity of crosslinks of the polymeric chains. In Sect. 2 we present the mathematical model for the drug transport using an integro-differential equation. An equivalent partial differential equation will be established. The solution of the differential problem whose spatial domain and differential equation change discretely in time is presented in Sect. 3. Some conclusions are presented in Sect. 4.

2 An Hybrid Non Fickian Mathematical Model

Let $\Omega(t),\ t \in [0, T_f]$ be the spatial domain that changes in time. We assume that $\Omega(t)$ is homogeneous and isotropic, we take $\Omega(t) = (-H(t), H(t))$. Considering that the concentration and temperature have symmetric profiles with respect to the origin, we take $\Omega(t) = [0, H(t)]$. At $x = 0$ we impose the symmetric boundary conditions and at $x = H(t)$ we assume that all the drug that attains the boundary is immediately removed. For $x \in \Omega(t)$, the drug concentration is described by the conservation equation

$$\frac{\partial c}{\partial t} = -\nabla(J_F(t) + J_{NF}(t)). \tag{5}$$

In (5), $J_F(t)$ and $J_{NF}(t)$ denote the Fickian and the non Fickian drug fluxes that are defined, respectively, by Fick's law (3) and by

$$J_{NF}(t) = -D_v(T)\nabla\sigma(t),$$

where $\sigma(t)$ represents the polymeric stress. As in [4, 5], we assume that the viscoelastic behavior of the polymer is described by the generalized Maxwell model also known as the Maxwell-Wiechert model with one arm

$$\sigma(t) = -\int_0^t \left(E_0 + E_1 e^{-\frac{E_1(t-s)}{\mu}}\right)\frac{\partial \varepsilon}{\partial s}ds, \tag{6}$$

where E_0, E_1 represent the Young modulus that depend on the temperature $T(t)$. In (6), μ denotes the viscosity of the polymer-solvent solution and ε represents the polymeric strain. We remark that a nonlinear relation between the strain ε and the concentration c was proposed for instance in [5]. This type of relation could be adopted in the model presented in this paper. However to simplify, we assume that $\varepsilon = \lambda c$. Then for the concentration we get

$$\begin{aligned}\frac{\partial c}{\partial t}(t) =& \nabla(D(t)\nabla c(t))\\ &- \lambda\nabla\left[D_v(t)\nabla\int_0^t\left(E_0(s) + E_1(s)e^{-\int_s^t \frac{E_1(\theta)}{\mu(\theta)}d\theta}\right)\frac{\partial c}{\partial s}(s)ds\right]\end{aligned} \tag{7}$$

in $(0, H(t)) \times (0, T_f]$, where $D(t) = D(T(t))$, $D_v(t) = D_v(T(t))$, $E_i(t) = E_i(T(t)), i = 0, 1$, and $\mu(t) = \mu(T(t))$. Equation (7) is completed with the boundary conditions

$$\nabla c(0, t) = 0,\ c(H(t), t) = 0,\ t \in (0, T_f]. \tag{8}$$

and the initial condition

$$c(0) = g \text{ in } (0, H(0)). \tag{9}$$

The IBVP (7), (8), (9) should be coupled with (1) for the temperature and a mathematical law for $H(t)$.

Following [11], we assume that the temperature switches between two values, below and above the critical solution temperature, leading to two different states for the spatial domain. Then $[0, T_f]$ is split into a number of subintervals, $[0, T_f] = \cup_{i=0}^{n-2}[t_i, t_{i+1}) \cup [t_{n-1}, t_n]$, $t_0 = 0$, $t_n = T_f$. We further assume that in the first time interval the polymeric structure is collapsed, which means that the temperature is above the critical temperature solution. Consequently, in $\cup_{i=0}[t_{2i}, t_{2i+1})$ the polymeric structure is in the dry state and in $\cup_{i=1}[t_{2i-1}, t_{2i})$ it is in the swollen state. Moreover, for $[t_j, t_{j+1})$ we have

$$\begin{aligned}\sigma_\ell(t) = -\lambda(E_{1,\ell} + E_{0,\ell})c_\ell(t) + \lambda\left(E_{0,\ell} + E_{1,\ell}e^{-\frac{E_{1,\ell}}{\mu_\ell}(t-t_j)}\right)c_\ell(t_j)\\ + \lambda\frac{E_{1,\ell}^2}{\mu_\ell}\int_{t_j}^t e^{-\frac{E_{1,\ell}}{\mu_\ell}(t-\theta)}c_\ell(\theta)d\theta,\end{aligned} \tag{10}$$

for $\ell = c, s$. The indices c and s are used to represent the dependence of the solutions and parameter values on the dry and swollen states, respectively. Taking in (7) the expression (10) we obtain to the following integro-differential equation

$$\begin{aligned}\frac{\partial c_\ell}{\partial t}(t) = (D - D_{v,\ell}\lambda\hat{E}_\ell)\Delta c_\ell(t) + D_{v,\ell}\lambda\frac{E_{1,\ell}^2}{\mu_\ell}\int_{t_j}^{t} e^{-\frac{E_{1,\ell}}{\mu_\ell}(t-\theta)}\Delta c_\ell(\theta)d\theta \\ + D_{v,\ell}\lambda\Big(E_{0,\ell} + E_{1,\ell}e^{-\frac{E_{1,\ell}}{\mu_\ell}(t-t_j)}\Big)\Delta c_\ell(t_j),\end{aligned} \tag{11}$$

where $\hat{E}_\ell = E_{0,\ell} + E_{1,\ell}$.

Finally, it is easy to show that c_ℓ satisfies

$$\frac{\partial^2 c_\ell}{\partial t^2} + \alpha_\ell\frac{\partial c_\ell}{\partial t} = D_{1,\ell}\Delta\frac{\partial c_\ell}{\partial t} + D_{2,\ell}\alpha_\ell\Delta c_\ell + \beta_\ell\alpha_\ell\Delta c_\ell(t_j) \text{ in } (0, H_\ell) \times (t_j, t_{j+1}), \tag{12}$$

where $\alpha_\ell = \frac{E_{1,\ell}}{\mu_\ell}$, $D_{1,\ell} = D - D_{v,\ell}\lambda\hat{E}_\ell$, $D_{2,\ell} = D - D_{v,\ell}\lambda E_{0,\ell}$, $\beta_\ell = D_{v,\ell}\lambda E_{0,\ell}$ and $\ell = c, s$. Equation (12) is complemented with the boundary conditions

$$\nabla c_\ell(0, t) = 0, c_\ell(H_\ell, t) = 0, \ t \in (t_j, t_{j+1}). \tag{13}$$

The main problem now is the definition of the initial conditions. It is clear that when $t \in [0, t_1)$, the initial conditions are given by

$$\begin{cases} c_c(0) = g \\ \dfrac{\partial c_c}{\partial t}(0) = D\Delta g \ \text{ in } (0, H_c). \end{cases} \tag{14}$$

To define the initial conditions for (12), from (11) we get

$$\frac{\partial c_\ell}{\partial t}(t_j) = D\Delta c_\ell(t_j) \ \text{ in } (0, H_\ell).$$

The question then arises. To solve the problem in the next time interval(t_1, t_2), what are the initial conditions that should be used in t_1? In the preceding step, that is in the computation of the solution in $[0, t1)$ we computed the value $c_\ell(t_1)$ in the spatial domain $[0, H(t_1))$. However this value can not be used as the initial condition for the next time step as the spatial domain is now $(0, H(t_2))$. More generally: What is the definition of $c_\ell(t_j)$? We follow in our approach the procedure presented in [11]. If in the interval (t_{j-1}, t_j) the polymeric structure is in the collapsed state, then a solution c_c defined in $[0, H_c] \times [t_{j-1}, t_j)$ was computed. However the initial conditions for (12) involve a function defined in $[0, H_s]$. One possibility to define the initial conditions for (12) is to extend c_c to $[0, H_s]$ constructing $c_{c,ext}$ such that

$$\int_0^{H_c} c_c(x, t_j)dx = \int_0^{H_s} c_{c,ext}(x)dx. \tag{15}$$

The idea underlying (15) is the conservation of mass: when the polymer swollens or shrinks the concentration changes but the mass is instantaneously constant. Then the initial conditions for (12) are defined by

$$\begin{cases} c_s(t_j) = c_{c,ext} \\ \dfrac{\partial c_s}{\partial t}(t_j) = D\Delta c_{c,ext} \ \text{ in } (0, H_c). \end{cases} \tag{16}$$

Summarizing, when the viscoelastic effect of the polymer is taken into account proposing the following algorithm to solve our problem. If the polymer is in the collapsed state at $t = 0$ then:

1. Solve the IBVP

$$\begin{cases} \dfrac{\partial^2 c_c}{\partial t^2} + \alpha_c \dfrac{\partial c_c}{\partial t} = D_{1,c}\Delta \dfrac{\partial c_c}{\partial t} + D_{2,c}\alpha_c \Delta c_c + \beta_c \alpha_c \Delta g \ \text{in}(0, H_c) \times (t_0, t_1], \\ \nabla c_c(0,t) = 0, c_c(H_c, t) = 0, \ t \in [t_0, t_1], \\ c_c(x, 0) = g(x), \ \dfrac{\partial c_c}{\partial t}(x, 0) = D\Delta g(x), \ \ x \in [0, H_c), \end{cases} \tag{17}$$

2. Extend $c_c(t_1)$ to $[0, H_s]$ constructing $c_{c,ext}$ such that

$$\int_0^{H_c} c_c(x, t_1) ds = \int_0^{H_s} c_{c,ext}(x) dx. \tag{18}$$

3. For $i = 1, \ldots, n-1$, solve the IBVP

$$\begin{cases} \dfrac{\partial^2 c_\ell}{\partial t^2} + \alpha_\ell \dfrac{\partial c_\ell}{\partial t} = D_{1,\ell}\Delta \dfrac{\partial c_\ell}{\partial t} + D_{2,\ell}\alpha_\ell \Delta c_\ell + \beta_\ell \alpha_\ell \Delta c_{ext}(t_i) \text{ in } (0, H_\ell) \times (t_i, t_{i+1}], \\ \nabla c_\ell(0,t) = 0, \ c_\ell(H_\ell, t) = 0, \ t \in [t_i, t_{i+1}], \\ c_\ell(x, t_i) = c_{ext}(x, t_i), \ \dfrac{\partial c_\ell}{\partial t}(x, t_i) = D \dfrac{\partial c_{ext}}{\partial t}(x, t_i), \ x \in [0, H_\ell] \end{cases} \tag{19}$$

with $\ell = c$ or $\ell = s$ for i even or odd, respectively, and c_{ext} is the extension of $c_\ell(t_i)$ defined in $[0, H^*]$ with $H^* = H_s$ or $H^* = H_c$ for i even or odd, respectively, satisfying

- if i is even
$$\int_0^{H_s} c_s(x, t_i) = \int_0^{H_c} c_{s,ext}(x) dx, \tag{20}$$

- if i is odd
$$\int_0^{H_c} c_c(x, t_i) = \int_0^{H_s} c_{c,ext}(x) dx. \tag{21}$$

3 An Analytic Solution

In this section, using Fourier analysis, we introduce the general explicit expressions for the solutions of the defined IBVP's. In the first result we establish a formal representation for the solution of the IBVP (17).

Theorem 1. *If $g \in L^2(\Omega)$ is such that $\nabla g \in L^2(\Omega)$ and $g(0) = \nabla g(H_c) = 0$, then*

$$\begin{aligned} c_c(x,t) = &\sum_{n \in I_{c,P}}^{\infty} cos\left(\frac{(2n+1)\pi}{2H_c}x\right)(A_n^0 e^{\omega_{+,c}t} + B_n^0 e^{\omega_{-,c}t}) \\ &+ \sum_{n \in I_{c,H}} cos\left(\frac{(2n+1)\pi}{2H_c}x\right) e^{Re_c t}(C_n^0 cos(\omega_c t) + D_n^0 sin(\omega_c t)) - \frac{\beta_c}{D_{2,c}} g(x) \end{aligned} \tag{22}$$

for $x \in [0, H_c], t \in [t_0, t_1]$, defines a formal solution $c_c(x,t)$ of the IBVP (22). In (22), $I_{c,P} = \{n \in \mathbb{N}_0 : n \geq n_+ \text{ or } n \leq n_-\}$, $I_{c,H} = \{n \in \mathbb{N}_0 : n_- < n < n_+\}$,

$$n_\pm = \left\lfloor \frac{1}{2}\left(\frac{2H_c\sqrt{\alpha_c}}{\pi}\frac{\sqrt{D_{2,c}} \pm \sqrt{D_{2,c} - D_{1,c}}}{D_{1,c}} - 1\right)\right\rfloor, \tag{23}$$

provided that $\sqrt{1 - \frac{D_{1,c}}{D_{2,c}}} + \sqrt{D_{1,c}}\sqrt{\frac{D_{1,c}}{D_{2,c}}}\frac{\pi}{2H_c\sqrt{\alpha_c}} < 1,$

$$\gamma_c = \left(\frac{(2n+1)\pi}{2H_c}\right)^2, \tag{24}$$

$$\omega_{\pm,c} = \frac{-(\alpha_c + \gamma_c D_{1,c}) \pm \sqrt{(\alpha_c + \gamma_c D_{1,c})^2 - 4\gamma_c \alpha_c D_{2,c}}}{2}, \tag{25}$$

$$Re_c = -\frac{\alpha_c + \gamma_c D_{1,c}}{2}, \tag{26}$$

$$\omega_c = \sqrt{-(\alpha_c + \gamma_c D_{1,c})^2 + 4\gamma_c \alpha_c D_{2,c}}, \tag{27}$$

and the Fourier coefficients $A_n^0, B_n^0, C_n^0, D_n^0$ are given by

$$A_n^0 = \frac{D_{2,c}\widehat{Dg''}(n) - \omega_{-,c}(D_{2,c} + \beta_c)\widehat{g}(n)}{D_{2,c}(\omega_{+,c} - \omega_{-,c})}, \tag{28}$$

$$B_n^0 = \frac{\omega_{+,c}(D_{2,c} + \beta_c)\widehat{g}(n) - D_{2,c}\widehat{Dg''}(n)}{D_{2,c}(\omega_{+,c} - \omega_{-,c})}, \tag{29}$$

$$C_n^0 = \frac{(D_{2,c} + \beta_c)\widehat{g}(n)}{D_{2,c}}, \tag{30}$$

and

$$D_n^0 = \frac{D_{2,c}D\widehat{g''}(n) - Re_c(D_{2,c} + \beta_c)\widehat{g}(n)}{D_{2,c}\omega_c} \tag{31}$$

where the notation $\hat{f}(n) = \dfrac{2}{H_c}\displaystyle\int_0^{H_c} f(x)cos\left(\frac{(2n+1)\pi}{2H_c}x\right)dx$ *was used.*

Proof. We start by the following convenient change of variable

$$c_c(x,t) = u_c(x,t) - \frac{\beta_c c_c(x,0)}{D_{2,c}} \tag{32}$$

that converts the nonhomogeneous IBVP (17) in the homogeneous one

$$\begin{cases} \dfrac{\partial u_c^2}{\partial t^2} + \alpha_c\dfrac{\partial u_c}{\partial t} = D_{1,c}\Delta\left(\dfrac{\partial u_c}{\partial t}\right) + D_{2,c}\alpha_c\Delta u_c(x,t),\ (x,t) \in (0,H_c)\times(t_0,t_1], \\ u_c(x,0) = \left(1 + \dfrac{\beta_c}{D_{2,c}}\right)g(x),\ \dfrac{\partial u_c}{\partial t}(x,0) = D\Delta g(x),\ x \in [0,H_c], \\ u_c(H_c,t) = 0,\ \nabla u_c(0,t) = 0,\ t \in [t_0,t_1]. \end{cases} \tag{33}$$

To obtain the solution of the new IBVP (33), we apply the method of separation of variables, defining $u_c(x,t) = X(x)T(t)$. Hence, replacing it in the partial differential equation of (33) we get

$$T^{''}(t)X(x) + \alpha_c X(x)T^{'}(t) = D_{1,c}X^{''}(x)T^{'}(t) + D_{2,c}\alpha_c X^{''}(x)T(t),$$

that leads to

$$\frac{T^{''}(t) + \alpha_c T^{'}(t)}{D_{1,c}T^{'}(t) + D_{2,c}\alpha_c T(t)} = \frac{X^{''}(x)}{X(x)} = -\gamma.$$

From the boundary conditions we obtain $X(H_c)T(t) = 0$ and $X^{'}(0)T(t) = 0$ and consequently we should have $X(H_c) = 0$ and $X'(0) = 0$. Then for X we obtain the boundary value problem

$$\begin{array}{l} X^{''}(x) + \gamma X(x) = 0,\ x \in (0,H_c), \\ X'(0) = 0, X(H_c) = 0, \end{array} \tag{34}$$

and for T we deduce

$$T^{''}(t) + (\alpha_c + \gamma D_{1,c})T^{'}(t) + \gamma D_{2,c}\alpha_c T(t) = 0.$$

We remark that if $\gamma \leq 0$, then $X(x) = 0$ that leads to the null solution. So, $\gamma > 0$, and

$$X(x) = A_n^0 cos(\sqrt{\gamma}x) + B_n^0 sen(\sqrt{\gamma}x).$$

As $X(H_c) = 0$ and $X^{'}(0) = 0$, we obtain

$$X(x) = cos\left(\frac{(2n+1)\pi}{2H_c}x\right), n \in \mathbb{N}_0,$$

and γ_c is given by (24).

On the other hand, to obtain T we notice that $z^2 + (\alpha_c + \gamma_c D_{1,c})z + \gamma_c\alpha_c D_{2,c} = 0$. Thus

$$z = \frac{-(\alpha_c + \gamma_c D_{1,c}) \pm \sqrt{(\alpha_c + \gamma_c D_{1,c})^2 - 4\gamma_c\alpha_c D_{2,c}}}{2}. \tag{35}$$

The definition of T depends on the nature of the roots defined by (35).

- If $(\alpha_c + \gamma_c D_1)^2 - 4\gamma_c\alpha_c D_{2,c} \geq 0$, (35) have the roots (25) and consequently, T is given by
$$T(t) = A_n^0 e^{\omega_{+,c}t} + B_n^0 e^{\omega_{-,c}t}. \tag{36}$$
- If $(\alpha_c + \gamma_c D_1)^2 - 4\gamma_c\alpha_c D_{2,c} < 0$, then
$$T(t) = \left(C_n^0 cos(\omega_c t) + D_n^0 sin(\omega_c t)\right)e^{Re_c t} \tag{37}$$
where Re_c and ω_c are given by (26) and (27).

To conclude the expression of u_c we need to specify the set of $n \in \mathbb{N}_0$ such that $(\alpha_c + \gamma_c D_1)^2 - 4\gamma_c\alpha_c D_{2,c} \geq 0$ or $(\alpha_c + \gamma_c D_1)^2 - 4\gamma_c\alpha_c D_{2,c} < 0$ holds. Let n_+ and n_- be defined by (41) which are the real zeros of $(\alpha_c + \gamma_c D_1)^2 - 4\gamma_c\alpha_c D_{2,c}$. Then, for $n \in I_{c,P} =]-\infty, n_-] \cup [n_+, +\infty[$, $T(t)$ is given by (36) and, for $n \in I_{c,H} =]n_-, n_+[$, $T(t)$ is given by (37). Consequently, the candidate to u_c admits the representation

$$\begin{aligned} u_c(x,t) = &\sum_{n \in I_{c,P}} cos\left(\frac{(2n+1)\pi}{2H_c}x\right)(A_n^0 e^{\omega_{+,c}t} + B_n^0 e^{\omega_{-,c}t}) \\ &+ \sum_{n \in I_{c,H}} cos\left(\frac{(2n+1)\pi}{2H_c}x\right)e^{Re_c t}(C_n^0 cos(\omega_c t) + D_n^0 sin(\omega_c)) \end{aligned}$$

where the constants A_n^0, B_n^0, C_n^0 and D_n^0 are computed using the initial conditions of the IBVP (33). Using the Fourier series of $\left(1 + \frac{\beta_c}{D_{2,c}}\right) g$ and $Dg^{''}$ we easily get the algebraic systems

$$A_n^0 + B_n^0 = \frac{(D_{2,c} + \beta_c)\widehat{g}(n)}{D_{2,c}}$$
$$\omega_{+,c}A_n^0 + \omega_{-,c}B_n^0 = \widehat{Dg^{''}}(n)$$

and

$$C_n^0 = \frac{(D_{2,c}+\beta_c)\widehat{g}(n)}{D_{2,c}}$$
$$Re_c C_n^0 + \omega_c D_n^0 = D\widehat{g''}(n),$$

where the notation $\hat{f}(n) = \frac{2}{H_c}\int_0^{H_c} f(x)cos\left(\frac{(2n+1)\pi}{2H_c}x\right)dx$ was used. Solving the last linear systems we get A_n^0, B_n^0, C_n^0, D_n^0 given by (28), (29), (30) and (31), respectively, that concludes the proof.

We observe that the computed solution is formal. To show that it is in fact solution of the IBVP (22) we need to prove that the series (22) defines a function c_c in $[0, H_c] \times [t_0, t_1]$ that is continuous, admits the partial derivatives that arise in the partial differential equation in (22) and satisfies all the conditions imposed to solve the problem.

To obtain a solution in the time interval $[t_1, t_2]$ we need to define an extension of $c_c(t_1)$, established in Theorem 1, to $[0, H_s]$ such that (18) holds. We start by noting that $c_c(x, t_1)$ can be rewritten in the following equivalent form

$$c_c(x,t_1) = \sum_{n=0}^{\infty} cos\left(\frac{(2n+1)\pi}{2H_c}x\right)C_n \tag{38}$$

where

$$C_n = \begin{cases} A_n^0 e^{\omega_{+,c}t_1} + B_n^0 e^{\omega_{-,c}t_1} - \frac{\beta_c}{D_{2,c}}\hat{g}(n) & n \in I_{c,P}, \\ e^{Re_c t_1}(C_n^0 cos(\omega_c t_1) + D_n^0 sin(\omega_c t_1)) - \frac{\beta_c}{D_{2,c}}\hat{g}(n) & n \in I_{c,H}, \end{cases}$$

with $\hat{g}(n) = \frac{2}{H_c}\int_0^{H_c} g(x)cos\left(\frac{(2n+1)\pi}{2H_c}x\right)dx$. We take

$$c_{c,ext}(x) = \frac{H_c}{H_s}\sum_{n=0}^{\infty} cos\left(\frac{(2n+1)\pi}{2H_s}x\right)C_n,\ x \in [0, H_s]. \tag{39}$$

The extension $c_{c,ext}$ defined by (39) satisfies (18) and its Fourier form is convenient to obtain easily the solution of the IBVP (19), with $\ell = s, i = 1$, in $[0, H_s] \times [t_1, t_2]$. In fact, applying Theorem 1

$$\begin{aligned} c_s(x,t) = &\sum_{n\in I_{s,P}}^{\infty} cos\left(\frac{(2n+1)\pi}{2H_s}x\right)(A_n^1 e^{\omega_{+,s}t} + B_n^1 e^{\omega_{-,s}t}) \\ &+ \sum_{n\in I_{s,H}} cos\left(\frac{(2n+1)\pi}{2H_s}x\right)e^{Re_s t}(C_n^1 cos(\omega_s t) + D_n^1 sin(\omega_s t)) - \frac{\beta_s}{D_{2,s}}c_{c,ext}(x,t_1) \end{aligned} \tag{40}$$

with $I_{s,P} = \{n \in \mathbb{N}_0 : n \geq n_+ \ or\ n \leq n_-\}$, $I_{s,H} = \{n \in \mathbb{N}_0 : n_- < n < n_+\}$,

$$n_{\pm} = \left\lfloor \frac{1}{2}\left(\frac{2H_s\sqrt{\alpha_s}}{\pi}\frac{\sqrt{D_{2,s}} \pm \sqrt{D_{2,s}-D_{1,s}}}{D_{1,s}} - 1\right)\right\rfloor, \tag{41}$$

provided that $\sqrt{1-\frac{D_{1,s}}{D_{2,s}}} + \sqrt{D_{1,s}}\sqrt{\frac{D_{1,s}}{D_{2,s}}}\frac{\pi}{2H_s\sqrt{\alpha_s}} < 1$,

$$\gamma_s = \left(\frac{(2n+1)\pi}{2H_s}\right)^2,$$

$$\omega_{\pm,s} = \frac{-(\alpha_s+\gamma_s D_{1,s}) \pm \sqrt{(\alpha_s+\gamma_s D_{1,s})^2 - 4\gamma_s\alpha_s D_{2,s}}}{2},$$

$$Re_s = -\frac{\alpha_s+\gamma_s D_{1,s}}{2},$$

$$\omega_s = \sqrt{-(\alpha_s+\gamma_s D_{1,s})^2 + 4\gamma_s\alpha_s D_{2,s}}, \tag{42}$$

and the Fourier coefficients $A_n^1, B_n^1, C_n^1, D_n^1$ are given by

$$A_n^1 = \frac{D_{2,s}D\widehat{c_{c,ext}(t_1)}''(n) - \omega_{-,s}(D_{2,s}+\beta_s)\widehat{c_{c,ext}(t_1)}(n)}{D_{2,s}(\omega_{+,s}-\omega_{-,s})},$$

$$B_n^1 = \frac{\omega_{+,s}(D_{2,s}+\beta_s)\widehat{c_{c,ext}(t_1)}(n) - D_{2,s}D\widehat{c_{c,ext}(t_1)}''(n)}{D_{2,s}(\omega_{+,s}-\omega_{-,s})},$$

$$C_n^1 = \frac{(D_{2,s}+\beta_s)\widehat{c_{c,ext}(t_1)}(n)}{D_{2,s}},$$

and

$$D_n^1 = \frac{D_{2,s}D\widehat{c_{c,ext}(t_1)}''(n) - Re_s(D_{2,s}+\beta_s)\widehat{c_{c,ext}(t_1)}(n)}{D_{2,s}\omega_s}$$

where $\widehat{c_{c,ext}(t_1)}(n) = \frac{2}{H_s}\int_0^{H_s} c_{c,ext}(x,t_1)cos\left(\frac{(2n+1)\pi}{2H_s}x\right)dx.$

To obtain the solution for $[0, H_c] \times [t_2, t_3]$ we apply again the Theorem 1 with the convenient adaptations.

4 Conclusions

The main objective of this work is the introduction of mathematical models for the drug release from polymeric thermoresponsive platforms. The polymer is a viscoelastic material where the Young modulus depends on temperature. The polymer has a lower critical solution temperature (LCST) and it switches from a collapsed state, for temperatures above the LCST, to a swollen state, for temperatures lower than the LCST.

To simulate the evolution of the polymeric platform, was assume that the change in the temperature leads to two different states of the polymeric structure. A hybrid model is obtained by splitting the time interval into disjoint subintervals, where a moving boundary polymeric domain is constructed. In each time subinterval, corresponding alternately, to a shrink and a swollen state, the drug transport is characterized by two sets of different values. An analytic approach based on Fourier analysis is proposed to construct the solution of the problem. The main theoretical result—Theorem 1, that allows the construction of a solution of the hybrid model, can be used to study its qualitative behaviour. Thermoresponsive polymers are attracting an enormous scientific interest for advanced applications in drug delivery. Mathematical modelling and simulation of drug delivery, from these materials, appears as an important co-adjutant in pioneering experimental studies. Though the work included in this paper still as an exploratory character, we think that promising numerical simulations can be obtained by using the Fourier approach presented here. In the near future we plan to develop this approach as well as the design of FEM/FDM well adapted to the moving boundary value problem.

Acknowledgements This work was supported by Centro de Matemática da Universidade de Coimbra UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MCTES and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020. The first and the second authors were also supported by the project *NEXT.parts—Next-generation of advanced hybrid parts*, funded by EU's Horizon 2020 science programme (Portugal 2020, COMPETE 2020).
Elisa Silveira was supported by the FCT PhD grant PD/BD/128058/2016 funded by the Portuguese Government through FCT/MCTES.

References

1. Alsuraifi, A., Curtis, A., Laamprou, D.: Stimuli responsive polymeric systems for cancer therapy. Pharmaceutis **10**, 136 (2018)
2. Cardoso, A., Calejo, M., Morais, C., Cardoso, L., Cruz, R., Zhu, K., Pedroso de Lima, M., Jurado, A., Nystr*ö*m, B.: Application of thermoresponsive PNIPAAM-b-PAMPTMA diblock copolymers in siRNA delivery. Mol. Pharm. **11**, 819–827 (2014)
3. Cattaneo, C.: Sulla conduzione del calore. Atti Semin. Mat. Fis. Univ. Modena Reggio Emilia **3**, 83–101 (1948)
4. Ferreira, J.A., Gudi*ñ*o, E., Grassi, M., OLiveira, P.: A 3D model for mechanistic control drug release. SIAM J. Appl. Math. **74**, 620–633 (2014)

5. Ferreira, J.A., Gudiño, E., Grassi, M., OLiveira, P.: A new look to non-Fickian diffusion. Appl. Math. Model. **39**, 194–204 (2015)
6. Gudiño, E.: Recent developments in non-Fickian diffusion: a new look at viscoelastic materials. Ph.D. thesis, University of Coimbra (2013)
7. Joseph, D., Preziosi, L.: Heat waves. Rev. Mod. Phys. **61**, 41–73 (1989)
8. Lanzalaco, S., Armelin, E.: Poly(N-isopropylacrylamide) and copolymers: a review on recent progresses in biomedical applications. Gels **3**, 36 (2017)
9. Rao, N., Ko, H., Lee, J., Park, J.: Recent progress and advances in stimuli-responsive polymers for cancer therapy. Front. Bioeng. Biotechnol. **6**, 110 (2018)
10. Toniolo, G., Efthimiadou, E., Kordas, G., Chatgilialoglu, K.: Development of multi-layered and multi-sensitive polymeric nanocontainers for cancer therapy: in vitro evaluation. Sci. Rep. **8**, 14704 (2018)
11. Tuoi, V.: Mathematical analysis of some models for drug delivery. Ph.D. thesis, School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway (2012)
12. Zhao, H., An, H., Xi, B., Yang, Y., Qin, J., Wang, Y., He, Y., Wang, X.: Self-heealing Hydrogels with both LCST and UCST through cross-linking induced thermo-response. Polymers **11**, 490 (2019)

Some Properties of a Generalized Solution for Shear Flow of a Compressible Viscous Micropolar Fluid Model

Loredana Simčić and Ivan Dražić

Abstract We consider the non-stationary shear flow of a compressible viscous and heat-conducting micropolar fluid between two parallel plates that present solid thermoinsulated walls, whereby the lower plate is fixed and the upper one is moving irrotationally. We assume that the fluid is perfect and polytropic in the thermodynamical sense, as well as that the initial density and temperature are strictly positive. We take smooth initial functions and analyze the corresponding problem with non-homogeneous boundary data for velocity and homogeneous boundary data for microrotation and heat flux.

In this work we give the overview of the current progress in mathematical analysis of the described problem with particular emphasis on the existence theorems and the uniqueness of the solution.

1 Introduction

The micropolar fluid model is a generalization of the classical Navier-Stokes model which allows the mathematical analysis of physical phenomena at the micro level. It was first introduced by Ahmed Cemal Eringen in the 1960s when he, besides classical hydrodynamical variables (such as mass density and velocity field), introduced the new vector field called microrotation velocity. It is important to note that in the micropolar fluid model the microdeformations are neglected.

In the last few years, the micropolar fluid model is widely applied in different engineering areas. For example, we can find the micropolar fluid model as the base model for liquid crystals with rigid molecules, magnetic fluids, clouds with dust, muddy

L. Simčić (✉) · I. Dražić
Faculty of Engineering, University of Rijeka, Rijeka, Croatia
e-mail: lsimcic@riteh.hr

I. Dražić
e-mail: idrazic@riteh.hr

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_35

fluids, some biological fluids, etc. [6]. For some specific applications, particularly in the field of medicine, we refer to [2], together with corresponding references.

Assumptions for the model in this paper are that the flow is isotropic, viscous and heat-conducting, as well as that the fluid is in the thermodynamical sense perfect and polytropic. Such a flow was first considered by Mujaković in [9]. Since then, the models with spherical and cylindrical symmetry of the solution were also analyzed [3, 7].

In this work we analyze the flow between two parallel thermo-insulated horizontal plates, with the upper one moving irrotationally. This kind of flow is called a shear flow and it has a great potential for applications, especially in lubrication theory, for example in the lubrication of magnetic disks, as it was stated in [5]. It is interesting to note that this model requires nonhomogeneous boundary conditions for velocity. For heat flux and microrotation field classical homogeneous boundary conditions are proposed.

The main goal of this paper is to give an overview of recent results concerning the shear flow model for compressible micropolar heat-conducting fluid. The paper is organized as follows: In the next section, we will derive the system which describes shear flow in the micropolar setting, first in the Eulerian, and then in the Lagrangian description. Then we will give an overview of the current progress in the mathematical analysis of this problem. We will introduce the generalized solution to the problem together with existence and uniqueness theorems.

2 The Mathematical Model

In this section we will describe the general mathematical model for the flow of the isotropic, viscous and heat-conducting compressible micropolar fluid stated, for example, in the book [8]. In the second part of the section, we present one-dimensional model derived from general model, which is used to describe the shear flow, as it was stated in [5]. The model is first given in Eulerian coordinates, and then in the Lagrangian.

The general model is:

$$\dot{\rho} = -\rho \nabla \cdot \mathbf{v}, \tag{1}$$

$$\rho \dot{\mathbf{v}} = \nabla \cdot \mathbf{T} + \rho \mathbf{f}, \tag{2}$$

$$\rho j_I \dot{\omega} = \nabla \cdot \mathbf{C} + \mathbf{T}_x + \rho \mathbf{g}, \tag{3}$$

$$\rho \dot{E} = -\nabla \cdot \mathbf{q} + \mathbf{T} : \nabla \mathbf{v} + \mathbf{C} : \nabla \omega - \mathbf{T}_x \cdot \omega, \tag{4}$$

$$\mathbf{T}_{ij} = (-p + \lambda \mathbf{v}_{k,k})\delta_{ij} + \mu \left(\mathbf{v}_{i,j} + \mathbf{v}_{j,i}\right) + \mu_r \left(\mathbf{v}_{j,i} - \mathbf{v}_{i,j}\right) - 2\mu_r \varepsilon_{mij} \omega_{\mathbf{m}}, \tag{5}$$

$$\mathbf{C}_{ij} = c_0 \omega_{\mathbf{k},\mathbf{k}} \delta_{\mathbf{ij}} + \mathbf{c_d} \left(\omega_{\mathbf{i},\mathbf{j}} + \omega_{\mathbf{j},\mathbf{i}}\right) + \mathbf{c_a} \left(\omega_{\mathbf{j},\mathbf{i}} - \omega_{\mathbf{i},\mathbf{j}}\right), \tag{6}$$

$$\mathbf{q} = -k_\theta \nabla\theta, \tag{7}$$

$$p = R\rho\theta, \tag{8}$$

$$E = c_v\theta, \tag{9}$$

defined on the domain $Q_T = \Omega \times]0, T[$, where $T > 0$ is arbitrary and $\Omega \subset \mathbf{R}^3$.

Here ρ, $\mathbf{v}$, ω, E, and θ are, respectively, the mass density, velocity, microrotation velocity, internal energy density and absolute temperature. $\mathbf{T}$ is the stress tensor, $\mathbf{C}$ is the couple stress tensor, $\mathbf{q}$ is the heat flux density vector, $\mathbf{f}$ is the body force density, $\mathbf{g}$ is the body couple density. p denotes pressure and the positive constant j_I is microinertia density. λ and μ are coefficients of viscosity and μ_r, c_0, c_d and c_a are coefficients of microviscosity. By the constant k_θ ($k_\theta \geq 0$) we denote the heat conduction coefficient. The positive constant R is the specific gas constant and the positive constant c_v denotes the specific heat at a constant volume.

Equations (1)–(4) are, respectively, local forms of the conservation laws for mass, momentum, momentum moment and energy. Equations (5)–(6) are constitutive equations for the micropolar continuum. Equation (7) is the Fourier law and Eqs. (8)–(9) present the assumptions that our fluid is perfect and polytropic. The coefficients of viscosity and the coefficients of microviscosity are related through the Clausius-Duhamel inequalities, as follows:

$$\mu \geq 0, \qquad 3\lambda + 2\mu \geq 0, \qquad \mu_r \geq 0. \tag{10}$$

$$c_d \geq 0, \quad 3c_0 + 2c_d \geq 0, \quad |c_d - c_a| \leq c_d + c_a. \tag{11}$$

Vector $\mathbf{T}_x$ in the Eqs. (3) and (4) is an axial vector with the Cartesian components $(\mathbf{T}_x)_i = \varepsilon_{ijk}\mathbf{T}_{jk}$. ε_{ijk} is the Levi-Civita symbol, δ_{ij} is Kronecker delta and we assume the Einstein notation for summation. The colon operator in Eq. (4) is the scalar product of tensors, for example $\mathbf{T} : \nabla\mathbf{v} = \mathbf{T}_{ji}\mathbf{v}_{i,j}$. The differential (dot) operator in Eqs. (1)–(4) denotes material derivative defined by

$$\dot{\mathbf{b}} = \partial_t\mathbf{b} + (\nabla\mathbf{b}) \cdot \mathbf{v}. \tag{12}$$

We analyze the flow between two parallel thermo-insulated horizontal plates, with the gravity force acting in the negative x-direction. We assume that lower plate is given by $x = 0$ and upper plate by $x = 1$. Because of these assumptions, we expect that the functions ρ, $\mathbf{v}$, ω and θ depend only on the vertical variable x and the time variable t, so we take

$$\rho(\mathbf{x}, t) = \rho(x, t), \quad \theta(\mathbf{x}, t) = \theta(x, t), \tag{13}$$

$$\mathbf{v}(\mathbf{x}, t) = \mathbf{v}(x, t) = \left(v^a(x, t), v^b(x, t), v^c(x, t)\right), \tag{14}$$

$$\omega(\mathbf{x},t) = \omega(x,t) = \left(\omega^a(x,t), \omega^b(x,t), \omega^c(x,t)\right), \tag{15}$$

for $(x,t) \in]0,1[\times]0,T[$. Let us note that in this way we still have three-dimensional fluid domain, but the corresponding problem became one-dimensional.

We assume that the lower plate is fixed and that the upper one is moving irrotationally. Therefore, we have the following boundary conditions for the velocity vector $\mathbf{v}$:

$$\mathbf{v}(0,t) = \mathbf{0}, \qquad \mathbf{v}(1,t) = \mathbf{a}(t), \tag{16}$$

where the vector $\mathbf{a}(t) = (0, a_1(t), a_2(t))$ defines the motion of the upper bounding plate. For the microrotation field and the heat flux we have standard homogeneous boundary conditions:

$$\omega(0,t) = \omega(1,t) = \mathbf{0}, \tag{17}$$

$$\partial_x\theta(0,t) = \partial_x\theta(1,t) = 0. \tag{18}$$

For initial conditions we take

$$\rho(x,0) = \rho_0(x), \quad \theta(x,0) = \theta_0(x), \tag{19}$$

$$\mathbf{v}(x,0) = \left(v_0^a(x), v_0^b(x), v_0^c(x)\right), \tag{20}$$

$$\omega(x,0) = \left(\omega_0^a(x), \omega_0^b(x), \omega_0^c(x)\right), \tag{21}$$

where the functions on the right-hand sides of the Eqs. (19)–(21) are smooth enough functions.

Now, using the assumptions (13)–(15), from (1)–(9) we obtain the model in the Eulerian description:

$$\partial_t\rho + v^a\,\partial_x\rho + \rho\,\partial_x v^a = 0, \tag{22}$$

$$\rho(\partial_t v^a + v^a\,\partial_x v^a) = -\partial_x(R\rho\theta) + (\lambda + 2\mu)\partial_{xx}v^a, \tag{23}$$

$$\rho(\partial_t v^b + v^a\,\partial_x v^b) = (\mu + \mu_r)\partial_{xx}v^b - 2\mu_r\partial_x\omega^c, \tag{24}$$

$$\rho(\partial_t v^c + v^a\,\partial_x v^c) = (\mu + \mu_r)\partial_{xx}v^c + 2\mu_r\partial_x\omega^b, \tag{25}$$

$$j_I\rho(\partial_t\omega^a + v^a\,\partial_x\omega^a) = (c_0 + 2c_d)\partial_{xx}\omega^a - 4\mu_r\omega^a, \tag{26}$$

$$j_I\rho(\partial_t\omega^b + v^a\,\partial_x\omega^b) = (c_d + c_a)\partial_{xx}\omega^b - 4\mu_r\omega^b - 2\mu_r\partial_x v^c, \tag{27}$$

$$j_I\rho(\partial_t\omega^c + v^a\,\partial_x\omega^c) = (c_d + c_a)\partial_{xx}\omega^c - 4\mu_r\omega^c + 2\mu_r\partial_x v^b, \tag{28}$$

$$\begin{aligned} c_v\rho(\partial_t\theta + v^a\,\partial_x\theta) &= k_\theta\partial_{xx}\theta - R\rho\theta\partial_x v^a \\ &+(\lambda+2\mu)(\partial_x v^a)^2 + (\mu+\mu_r)\big((\partial_x v^b)^2 + (\partial_x v^c)^2\big) \\ &+4\mu_r\big((\omega^a)^2 + (\omega^b)^2 + (\omega^c)^2\big) + 4\mu_r(\omega^b\partial_x v^c - \omega^c\partial_x v^b) \\ &+(c_0+2c_d)(\partial_x\omega^a)^2 + (c_d+c_a)\big((\partial_x\omega^b)^2 + (\partial_x\omega^c)^2\big), \end{aligned} \tag{29}$$

and then in the Lagrangian description:

$$\partial_t\rho + \rho^2\partial_x v^a = 0, \tag{30}$$

$$\partial_t v^a = -\partial_x(R\rho\theta) + (\lambda+2\mu)\partial_x(\rho\partial_x v^a), \tag{31}$$

$$\partial_t v^b = (\mu+\mu_r)\partial_x(\rho\partial_x v^b) - 2\mu_r\partial_x\omega^c, \tag{32}$$

$$\partial_t v^c = (\mu+\mu_r)\partial_x(\rho\partial_x v^c) + 2\mu_r\partial_x\omega^b, \tag{33}$$

$$j_I\partial_t\omega^a = (c_0+2c_d)\partial_x(\rho\partial_x\omega^a) - 4\mu_r\frac{\omega^a}{\rho}, \tag{34}$$

$$j_I\partial_t\omega^b = (c_d+c_a)\partial_x(\rho\partial_x\omega^b) - 4\mu_r\frac{\omega^b}{\rho} - 2\mu_r\partial_x v^c, \tag{35}$$

$$j_I\partial_t\omega^c = (c_d+c_a)\partial_x(\rho\partial_x\omega^c) - 4\mu_r\frac{\omega^c}{\rho} + 2\mu_r\partial_x v^b, \tag{36}$$

$$\begin{aligned} c_v\partial_t\theta &= k_\theta\partial_x(\rho\partial_x\theta) - R\rho\theta\,\partial_x\omega^a + (\lambda{+}2\mu)\rho(\partial_x v^a)^2 + (\mu\,{+}\mu_r)\rho((\partial_x v^b)^2 \\ &+(\partial_x v^c)^2) + 4\mu_r\frac{1}{\rho}((\omega^a)^2 + (\omega^b)^2 + (\omega^c)^2) + 4\mu_r(\omega^b\partial_x v^c - \omega^c\partial_x v^b) \\ &+(c_0+2c_d)\rho(\partial_x\omega^a)^2 + (c_d+c_a)\rho((\partial_x\omega^b)^2 + (\partial_x\omega^c)^2). \end{aligned} \tag{37}$$

As it was shown in [5], in order to homogenize the boundary condition for $\mathbf{v}(1,t)$, we take the substitution

$$\mathbf{V}(x,t) = \mathbf{u}(x,t) - \mathbf{h}(x,t), \tag{38}$$

where

$$\mathbf{u} = (0, v^b, v^c), \tag{39}$$

and

$$\mathbf{h}(x,t) = \mathbf{a}(t)\int_0^x \frac{dy}{\rho(y,t)}. \tag{40}$$

Finally, we obtain the system

$$\partial_t \rho + \rho^2 \partial_x v^a = 0, \tag{41}$$

$$\partial_t v^a = -\partial_x (R\rho\theta) + (\lambda + 2\mu)\partial_x(\rho\, \partial_x v^a), \tag{42}$$

$$\partial_t \mathbf{V} = (\mu + \mu_r)\partial_x(\rho\, \partial_x \mathbf{V}) + 2\mu_r \nabla \times \mathbf{w} - \partial_t \mathbf{h}, \tag{43}$$

$$j_I \partial_t \omega^a = (c_0 + 2c_d)\partial_x(\rho\, \partial_x \omega^a) - 4\mu_r \frac{\omega^a}{\rho}, \tag{44}$$

$$j_I \partial_t \mathbf{w} = (c_d + c_a)\partial_x(\rho\, \partial_x \mathbf{w}) - \frac{4\mu_r}{\rho}\mathbf{w} + 2\mu_r \nabla \times \mathbf{V} + 2\mu_r \nabla \times \mathbf{h}, \tag{45}$$

$$\begin{aligned} c_v \partial_t \theta &= k_\theta \partial_x(\rho\, \partial_x \theta) - R\rho\theta\partial_x v^a + (\lambda + 2\mu)\rho(\partial_x v^a)^2 + (\mu + \mu_r)\rho|\partial_x \mathbf{V}|^2 \\ &+2(\mu + \mu_r)\rho\partial_x \mathbf{V}\partial_x \mathbf{h} + (\mu + \mu_r)\rho|\partial_x \mathbf{h}|^2 + 4\mu_r \frac{|\mathbf{w}|^2}{\rho} - 4\mu_r(\nabla \times \mathbf{V}) \cdot \mathbf{w} \\ &+4\mu_r \frac{w^2}{\rho} - 4\mu_r(\nabla \times \mathbf{h}) \cdot \mathbf{w} + (c_0 + 2c_d)\rho(\partial_x \omega^a)^2 + (c_d + c_a)\rho|\partial_x \mathbf{w}|^2, \end{aligned} \tag{46}$$

$$v^a(0,t) = v^a(1,t) = 0, \quad \mathbf{V}(0,t) = \mathbf{V}(1,t) = 0, \tag{47}$$

$$\omega^a(0,t) = \omega^a(1,t) = 0, \quad \mathbf{w}(0,t) = \mathbf{w}(1,t) = 0, \tag{48}$$

$$\partial_x \theta(0,t) = \partial_x \theta(1,t) = 0, \tag{49}$$

$$\rho(x,0) = \rho_0(x), \quad \theta(x,0) = \theta_0(x), \tag{50}$$

$$v^a(x,0) = v_0^a(x), \quad \mathbf{V}(x,0) = \mathbf{V}_0(x) = \mathbf{u}_0(x) - \mathbf{a}(0)\int_0^x \frac{dy}{\rho_0(y)}, \tag{51}$$

$$\omega^a(x,0) = \omega_0^a(x), \quad \mathbf{w}(x,0) = \mathbf{w}_0(x), \tag{52}$$

where

$$\mathbf{w} = (0, \omega^b, \omega^c). \tag{53}$$

3 Properties of the Solution

In this section we consider the properties of the so-called generalized solution to the problem (41)–(52), defined for example in [10]. For the readers' convenience, we state it here also.

Definition 1. A generalized solution of the problem (41)–(52) in the domain $Q_T =]0,1[\times]0,T[$ is a function

$$(x,t) \mapsto (\rho, v^a, v^b, v^c, \omega^a, \omega^b, \omega^c, \theta)(x,t), \quad (x,t) \in Q_T, \tag{54}$$

where

$$\rho \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(]0,1[)) \cap \mathrm{H}^1(Q_T)\ ,\ \inf_{Q_T} \rho > 0\ , \tag{55}$$

$$v^a, v^b, v^c, \omega^a, \omega^b, \omega^c, \theta \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(]0,1[)) \cap \mathrm{H}^1(Q_T) \cap \mathrm{L}^2(0,T;\mathrm{H}^2(]0,1[)), \tag{56}$$

that satisfies Eqs. (41)–(46) a.e. in Q_T and conditions (47)–(52) in the sense of traces.

The existence of the generalized solution to the problem (41)–(52) was analyzed first. Using the Faedo–Galerkin method, Ivan Dražić proved in [4] the existence locally in time. The result is summarized in the following theorem:

Theorem 1. *Let the initial functions satisfy the conditions*

$$\rho_0(x) \geq m, \quad \theta_0(x) \geq m \quad \textit{for} \ \ x \in]0,1[, \tag{57}$$

where $m \in \mathbf{R}^+$, *as well as*

$$\rho_0, \theta_0 \in \mathrm{H}^1(]0,1[), \quad v_0^a, v_0^b, v_0^c, \omega_0^a, \omega_0^b, \omega_0^c \in \mathrm{H}_0^1(]0,1[), \tag{58}$$

$$a_1(t), a_2(t) \in \mathrm{H}^2(]0,T[). \tag{59}$$

Then there exists T_0, $0 < T_0 \leq T$, *such that the desribed problem has a generalized solution in* $Q_0 = Q_{T_0}$, *having the property*

$$\theta > 0 \ \ \textit{in} \ \overline{Q}_0. \tag{60}$$

Now, as we know that the solution exists, we can start to analyze the properties of the solution. The first property which was analysed is the uniqueness of a generalized solution, which was proved in [10]. The result is stated in the next theorem.

Theorem 2. *If the problem (41)–(52) has a generalized solution defined by (54)–(56), then this solution is unique.*

To prove Theorem 2, the method based on forming an auxiliary system for the difference of two solutions and analysis of its solution was used. This method was described in [1], where it has been applied for the one-dimensional case of a classical fluid. In what follows, we will give a sketch of the proof of Theorem 2.

3.1 Sketch of the Proof of Theorem 2

For simplicity, as given in [10], hereafter we consider the specific volume $u = \rho^{-1}$ instead of the density ρ. Now we assume that

$$(u_i, v_i^a, v_i^b, v_i^c, \omega_i^a, \omega_i^b, \omega_i^c, \theta_i), \quad i = 1, 2 \tag{61}$$

are two distinct generalized solutions of the problem (41)–(52) in the domain Q_T with the properties (57)–(58), (60).

We introduce the functions $u = u_1 - u_2$, $v^a = v_1^a - v_2^a$, $v^b = v_1^b - v_2^b$, $v^c = v_1^c - v_2^c$, $\omega^a = \omega_1^a - \omega_2^a$, $\omega^b = \omega_1^b - \omega_2^b$, $\omega^c = \omega_1^c - \omega_2^c$ and $\theta = \theta_1 - \theta_2$.

After some calculations it can be shown that $(u, v^a, v^b, v^c, \omega^a, \omega^b, \omega^c, \theta)$ satisfy the following system:

$$\partial_t u = \partial_x v^a, \tag{62}$$

$$\partial_t v^a = -R\partial_x\left(\frac{\theta}{u_1}\right) + R\partial_x\left(\frac{\theta_2 u}{u_1 u_2}\right) + (\lambda + 2\mu)\partial_x\left(\frac{1}{u_1}\partial_x v^a - \frac{u}{u_1 u_2}\partial_x v_2^a\right), \tag{63}$$

$$\partial_t v^b = (\mu + \mu_r)\partial_x\left(\frac{1}{u_1}\partial_x v^b - \frac{u}{u_1 u_2}\partial_x v_2^b\right) - 2\mu_r \partial_x \omega^c, \tag{64}$$

$$\partial_t v^c = (\mu + \mu_r)\partial_x\left(\frac{1}{u_1}\partial_x v^c - \frac{u}{u_1 u_2}\partial_x v_2^c\right) + 2\mu_r \partial_x \omega^b, \tag{65}$$

$$j_I \partial_t \omega^a = (c_0 + 2c_d)\partial_x\left(\frac{1}{u_1}\partial_x \omega^a - \frac{u}{u_1 u_2}\partial_x \omega_2^a\right) - 4\mu_r\left(u_1\omega^a + u\omega_2^a\right), \tag{66}$$

$$j_I \partial_t \omega^b = (c_d + c_a)\partial_x\left(\frac{1}{u_1}\partial_x \omega^b - \frac{u}{u_1 u_2}\partial_x \omega_2^b\right) - 4\mu_r\left(u_1\omega^b + u\omega_2^b\right) - 2\mu_r \partial_x v^c, \tag{67}$$

$$j_I \partial_t \omega^c = (c_d + c_a)\partial_x\left(\frac{1}{u_1}\partial_x \omega^c - \frac{u}{u_1 u_2}\partial_x \omega_2^c\right) - 4\mu_r\left(u_1\omega^c + u\omega_2^c\right) + 2\mu_r \partial_x v^b, \tag{68}$$

$$
\begin{aligned}
c_v \partial_t \theta = k\partial_x \left(\frac{1}{u_1}\partial_x\theta - \frac{u}{u_1 u_2}\partial_x\theta_2\right) - R\frac{1}{u_1}\left(\theta_1\partial_x\omega^a + \theta\partial_x\omega_2^a\right) \\
+R\frac{u}{u_1u_2}\theta_2\partial_x\omega_2^a + (\lambda+2\mu)\frac{1}{u_1}\partial_x v^a\left(\partial_x v_1^a + \partial_x v_2^a\right) - (\lambda+2\mu)\frac{u}{u_1u_2}(\partial_x v_2^a)^2 \\
+(\mu+\mu_r)\frac{1}{u_1}\partial_x v^b\left(\partial_x v_1^b + \partial_x v_2^b\right) - (\mu+\mu_r)\frac{u}{u_1u_2}(\partial_x v_2^b)^2 \\
+(\mu+\mu_r)\frac{1}{u_1}\partial_x v^c\left(\partial_x v_1^c + \partial_x v_2^c\right) - (\mu+\mu_r)\frac{u}{u_1u_2}(\partial_x v_2^c)^2 \\
+4\mu_r u_1\left(\omega^a\left(\omega_1^a+\omega_2^a\right) + \omega^b\left(\omega_1^b+\omega_2^b\right) + \omega^c\left(\omega_1^c+\omega_2^c\right)\right) \\
+4\mu_r u\left((\omega_2^a)^2 + (\omega_2^b)^2 + (\omega_2^c)^2\right) \\
+4\mu_r\left(\omega_1^b\partial_x v^c + \omega^b\partial_x v_2^c - \omega_1^c\partial_x v^b - \omega^c\partial_x v_2^b\right) \\
+(c_0+2c_d)\frac{1}{u_1}\partial_x\omega^a\left(\partial_x\omega_1^a + \partial_x\omega_2^a\right) - (c_0+2c_d)\frac{u}{u_1u_2}(\partial_x\omega_2^a)^2 \\
+(c_d+c_a)\frac{1}{u_1}\partial_x\omega^b\left(\partial_x\omega_1^b + \partial_x\omega_2^b\right) - (c_d+c_a)\frac{u}{u_1u_2}(\partial_x\omega_2^b)^2 \\
+(c_d+c_a)\frac{1}{u_1}\partial_x\omega^c\left(\partial_x\omega_1^c + \partial_x\omega_2^c\right) - (c_d+c_a)\frac{u}{u_1u_2}(\partial_x\omega_2^c)^2, \qquad (69)
\end{aligned}
$$

with initial and boundary conditions for the functions u, v^a, v^b, v^c, ω^a, ω^b, ω^c and θ:

$$
\begin{aligned}
u(x,0)=0, \quad v^a(x,0)=0, \quad v^b(x,0)=0, \quad v^c(x,0)=0, \\
\omega^a(x,0)=0, \quad \omega^b(x,0)=0, \quad \omega^c(x,0)=0, \quad \theta(x,0)=0, \qquad (70)
\end{aligned}
$$

$$
\begin{aligned}
v^a(0,t)=v^a(1,t)=0, \quad v^b(0,t)=v^b(1,t)=0, \ v^c(0,t)=v^c(1,t)=0, \\
\omega^a(0,t)=\omega^a(1,t)=0, \ \omega^b(0,t)=\omega^b(1,t)=0, \ \omega^c(0,t)=\omega^c(1,t)=0, \\
\partial_x\theta(0,t)=\partial_x\theta(1,t)=0 \qquad (71)
\end{aligned}
$$

for $x \in]0,1[$, $t \in]0,T[$.

Next, in [10] a series of estimates for the solution of the problem (62)–(71) were derived, which we summarize in the following lemma:

Lemma 1. *There exists a constant $C > 0$ such that for any $t \in]0,T[$ we have:*

$$\left\|u(t)\right\|^2 \le C\int_0^t \left\|\partial_x v^a(\tau)\right\|^2 d\tau. \qquad (72)$$

$$\left\|v^a(t)\right\|^2 + \int_0^t \left\|\partial_x v^a(\tau)\right\|^2 d\tau \le C\int_0^t \left\|\theta(\tau)\right\|^2 d\tau. \qquad (73)$$

$$\left\|v^b(t)\right\|^2 + \int_0^t \left\|\partial_x v^b(\tau)\right\|^2 d\tau \le C\int_0^t \left(\left\|\theta(\tau)\right\|^2 + \left\|\partial_x\omega^c(\tau)\right\|^2\right) d\tau, \qquad (74)$$

$$\left\|v^c(t)\right\|^2+\int_0^t\left\|\partial_x v^c(\tau)\right\|^2 d\tau \le C\int_0^t\left(\left\|\theta(\tau)\right\|^2+\left\|\partial_x\omega^b(\tau)\right\|^2\right)d\tau. \tag{75}$$

$$\left\|\omega^a(t)\right\|^2+\int_0^t\left\|\partial_x\omega^a(\tau)\right\|^2 d\tau \le C\int_0^t\left\|\theta(\tau)\right\|^2 d\tau, \tag{76}$$

$$\left\|\omega^b(t)\right\|^2+\int_0^t\left\|\partial_x\omega^b(\tau)\right\|^2 d\tau \le C\int_0^t\left\|\theta(\tau)\right\|^2 d\tau+\varepsilon\int_0^t\left\|\partial_x v^c(\tau)\right\|^2 d\tau, \tag{77}$$

$$\left\|\omega^c(t)\right\|^2+\int_0^t\left\|\partial_x\omega^c(\tau)\right\|^2 d\tau \le C\int_0^t\left\|\theta(\tau)\right\|^2 d\tau+\varepsilon\int_0^t\left\|\partial_x v^b(\tau)\right\|^2 d\tau, \tag{78}$$

where $\varepsilon > 0$ is arbitrary.

For the proof of Lemma 1 we use (62)–(68) and properties of the generalized solution, Hölder, Young and Gronwall inequalities, as well as inequalities

$$|f|^2 \le 2\left\|f\right\|\left\|\partial_x f\right\|,\quad \left\|f\right\| \le 2\left\|\partial_x f\right\|, \tag{79}$$

$$|\partial_x f|^2 \le 2\left\|\partial_x f\right\|\left\|\partial_{xx} f\right\|,\quad \left\|\partial_x f\right\| \le 2\left\|\partial_{xx} f\right\|, \tag{80}$$

which are derived from the well-known Friedrichs, Poincare and Gagliardo-Ladyzhenskaya inequalities.

By combining inequalities obtained in Lemma 1, we obtain the following lemma:

Lemma 2. *There exists a constant $C > 0$ such that for any $t \in]0, T[$ we have*

$$\begin{aligned}&\left\|u(t)\right\|^2+\left\|v^a(t)\right\|^2+\left\|v^b(t)\right\|^2+\left\|v^c(t)\right\|^2+\left\|\omega^a(t)\right\|^2+\left\|\omega^b(t)\right\|^2\\&\quad+\left\|\omega^c(t)\right\|^2+\int_0^t\left(\left\|\partial_x v^a(\tau)\right\|^2+\left\|\partial_x v^b(\tau)\right\|^2+\left\|\partial_x v^c(\tau)\right\|^2\right.\\&\left.+\left\|\partial_x\omega^a(\tau)\right\|^2+\left\|\partial_x\omega^b(\tau)\right\|^2+\left\|\partial_x\omega^c(\tau)\right\|^2\right)d\tau \le C\int_0^t\left\|\theta(\tau)\right\|^2 d\tau.\end{aligned} \tag{81}$$

In a similar way as in Lemma 1, in [10] we get

Lemma 3. *There exists a constant $C > 0$ such that for any $t \in]0, T[$ we have*

$$\left\|\theta(t)\right\|^2+\int_0^t\left\|\partial_x\theta(\tau)\right\|^2 d\tau \le C\int_0^t\left\|\theta(\tau)\right\|^2 d\tau. \tag{82}$$

By applying Gronwall inequality to (82), we get $\theta = 0$ on Q_T. From (81) we get

$$u=0,\quad v^a=v^b=v^c=0\quad \omega^a=\omega^b=\omega^c=0\quad on\quad Q_T, \tag{83}$$

which proves Theorem 2.

4 Conclusion

In this paper, the overview of the recent results for the shear flow of a compressible micropolar fluid flow was given. As it was stated before, we know that corresponding non-homogeneous boundary condition problem has a unique local solution. Let us also note that the numerical solution for this problem is also constructed in [5], whereby two different methods were used; the first one is the Faedo-Galerkin approximations, and the second one is the finite difference method. The other properties, such as global existence, stabilization, as well as regularity of the solution, are currently subjects of our research.

Acknowledgements This work has been fully supported by the University of Rijeka, Croatia under the project 17.10.2.2.01 (Initial-boundary problems in the research of the micropolar continuum model) and the project uniri-prirod-18-118-1257 (Analysis of mathematical models of fluid mechanics and technical systems using data-driven algorithms for Koopman operator).

References

1. Antontsev, S.N., Kazhikhov, A.V., Monakhov, V.N.: Boundary Value Problems in Mechanics of Nonhomogeneous Fluid. Elsevier, North Holland (1990)
2. Dražić, I.: Dimensionless formulation for the one-dimensional compressible flow of the viscous and heat-conducting micropolar fluid. Phys. Astron. Int. J. **2**(5), 420–423 (2018)
3. Dražić, I.: 3-D flow of a compressible viscous micropolar fluid model with spherical symmetry: a brief survey and recent progress. Rev. Math. Phys. **30**, 1830001 (2018)
4. Dražić, I.: A shear flow problem for compressible viscous and heat conducting micropolar fluid: local existence theorem (preprint)
5. Dražić, I., Črnjarić-Žic, N., Simčić, L.: A shear flow problem for compressible viscous micropolar fluid: derivation of the model and numerical solution. Math. Comput. Simul. **162**, 249–267 (2019)
6. Eringen, A.C.: Microcontinuum Field Theories II: Fluent Media. Springer, New York (2001)
7. Huang, L., Dražić, I.: Large-time behavior of solutions to the 3-D flow of a compressible viscous micropolar fluid with cylindrical symmetry. Math. Methods Appl. Sci. (2018). https://doi.org/10.1002/mma.5250
8. Lukaszewicz, G.: Micropolar Fluids: Theory and Applications. Modeling and Simulation in Science, Engineering and Technology, Birkhäuser, Boston (1999)
9. Mujaković, N.: One-dimensional flow of a compressible viscous micropolar fluid: a local existence theorem. Glas. Mat. Ser. III **33**(1), 71–91 (1998)
10. Simčić, L.: A shear flow problem for compressible viscous micropolar fluid: uniqueness of a generalized solution. Math. Meth. Appl. Sci. **42**, 6358–6368 (2019). https://doi.org/10.1002/mma.5727

Collocation Solution of Fractional Differential Equations in Piecewise Nonpolynomial Spaces

M. Luísa Morgado and Magda Rebelo

Abstract In this paper, we develop an accurate collocation scheme for ordinary initial value problems of the Caputo type and in order to illustrate the performance of the method we provide several numerical examples. At the end, we also indicate how this method can be used to approximate the solution of time-fractional diffusion equations. At it will be seen, it allows us to obtain accurate solutions even when the solution is not smooth at the origin.

1 Introduction

Our concern here is the accurate numerical solution of initial value problems (IVPs) for ordinary fractional differential equations of the Caputo type:

$$D^{\alpha} y(t) = f(t, y(t)),\ \ 0 < t \leq T, \tag{1}$$

$$y(0) = y_0, \tag{2}$$

where α is a real number such that $0 < \alpha < 1$, $T > 0$ and D^{α} denotes the Caputo differential operator of order $\alpha \notin \mathbb{N}$, defined by [3]:

$$D^{\alpha} y(t) := {}^{RL}D^{\alpha}(y - T[y])(t),$$

M. L. Morgado (✉)
CEMAT, Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

Departamento de Matematica, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
e-mail: luisam@utad.pt

M. Rebelo (✉)
Centro de Matemática e Aplicações (CMA) and Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Caparica, Portugal
e-mail: msjr@fct.unl.pt

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_36

where $T[y]$ is the Taylor polynomial of degree $\lfloor \alpha \rfloor$ for y, centered at 0, and ${}^{RL}D^{\alpha}$ is the Riemann-Liouville derivative of order α [7].

The latter is defined by ${}^{RL}D^{\alpha} := D^{\lceil \alpha \rceil} J^{\lceil \alpha \rceil - \alpha}$, with J^{β} being the Riemann-Liouville integral operator,

$$J^{\beta} y(t) := \frac{1}{\Gamma(\beta)} \int_0^t (t-s)^{\beta-1} y(s) ds$$

and $D^{\lceil \alpha \rceil}$ is the classical integer order derivative, where $\lceil \alpha \rceil$ is the smallest integer greater than or equal to α.

We also assume that the right hand-side function satisfies a Lipschitz condition with respect to the second variable:

$$|f(t, y) - f(t, z)| \leq L\, |y - z|\,, \tag{3}$$

for some constant $L > 0$ independent of t, y and z, which is a necessary condition to ensure the uniqueness of the solution (see [4]).

In the last decades, a huge amount of papers addressed this subject, by reporting a series of numerical schemes for the solution of (1)–(2). We refer the book of Li and Zeng [5] for a survey of those works. Most of the developed numerical schemes are based in finite differences formulas, assuming a certain smoothness of the solution. Whenever this is not the case, it will be convenient to use graded meshes reflecting the singularities of the solution, in order to recover the optimal orders of convergence (see [8]).

Opposed to the integer-order case, fractional-order problems as (1)–(2) may possess nonsmooth solutions even when the data is sufficiently smooth.

In [4] the following result, analysing the behavior of the solution near the origin, was proved:

Lemma 1 ([4]). *Assume that the solution of (1)–(2) exists and is unique on* $[0, T]$*, for a certain* $T > 0$*. If* $\alpha = \frac{p}{q}$*, where* $p \geq 1$ *and* $q \geq 2$ *are two relatively prime integers and if the right-hand side function* f *can be written in the form* $f(t, y(t)) = \overline{f}(t^{1/q}, y(t))$*, where* $\overline{f}$ *is analytic in a neighborhood of* $(0, y_0)$*, then there exists a* $r > 0$ *such that the unique solution of the problem (1)–(2) can be represented in terms of powers of* $t^{1/q}$*:*

$$y(t) = \sum_{i=0}^{\infty} a_i t^{i/q}, t \in [0, r), \tag{4}$$

where the a_i *are constants.*

Results on the existence and uniqueness of solution of such problems may be found in [4] as well their equivalence to Volterra integral equations with weakly singular kernels:

Lemma 2 ([4]). *If the function f is continuous, the initial value problem (1)–(2) is equivalent to the following Volterra integral equation of the second kind:*

$$y(t) = y_0 + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, y(s)) ds. \tag{5}$$

In [6] a nonpolynomial collocation method was derived for problems of the form (1)–(2) taking into account their equivalent integral representation (2). Choosing suitable piecewise nonpolynomial spaces, the method there was able to deal with the same accuracy for problems with smooth and nonsmooth solutions.

Here, we will use a different approach. We will find the collocation solution of (1)–(2) directly without falling back on (5). The obtained numerical scheme is easier to derive and implement than the one obtained in [6] and, as we will see in the forthcoming section, it attains the same accuracy than the one in [6].

2 Numerical Scheme

For $\alpha \in (0, 1)$, taking the definition of the Caputo derivative into account, (1) may be rewritten as:

$$\frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_0^t (t-s)^{-\alpha} \left(y(s) - y_0\right) \, ds = f(t, y(t)), \; t > 0. \tag{6}$$

Let us consider a uniform partition of $I = [0, T]$

$$I_h = \{0 = t_0 < t_1 < \ldots < t_N = T\},$$

where $t_i = ih$, $i = 0, \ldots, N$, $h = \frac{T}{N}$, and set

$$\sigma_0 = [0, t_1], \;\; \sigma_n = (t_n, t_{n+1}], \;\; n = 1, \ldots, N-1. \tag{7}$$

In order to deal with the potential singularities of the solution (cf. (4)), given $m \in \mathbb{N}$, we define the nonpolynomial space

$$V_m^\alpha = span\left\{t^{i+j\alpha}, \quad i, j \in \mathbb{N}_0, \; i + j\alpha < m\right\}. \tag{8}$$

In order to simplify the above notation we introduce the index set

$$W_{\alpha,m} = \{i + j\alpha, \quad i, j \in \mathbb{N}_0, \; i + j\alpha < m\} = \{\nu_k : k = 1, ..., \ell\}.$$

Using this notation we can write V_m^α as

$$V_m^\alpha = span\{t^{\nu_k}, \; k = 1, ..., \ell\}. \tag{9}$$

Let

$$X_h = \{t_{ni} = t_n + c_i h : \ 0 \le c_1 < c_2 < \ldots < c_\ell \le 1, \ n = 0, 1, \ldots, N-1\} \quad (10)$$

be a set of collocation points, determined by the given mesh I_h and the collocation parameters c_i, $i = 1, \ldots, \ell$.
Defining the piecewise nonpolynomial spaces:

$$V^{\alpha}_{h,m} = \left\{ v : \ v|_{\sigma_i} \in V^{\alpha}_{m}, \ i = 0, 1, \ldots, N-1 \right\},$$

the collocation solution $v \in V^{\alpha}_{h,m}$ will be found by requiring that v satisfies the given fractional differential equation (1) on a subset X_h and coincides with exact solution $y(t)$ at $t = 0$. That is,

$$v(0) = y_0, \quad (11)$$

$$\left[\frac{d}{dt} \left(\sum_{j=0}^{i-1} \int_{t_j}^{t_{j+1}} (t-s)^{-\alpha} \left(v(s) - y_0\right) \, ds + \int_{t_i}^{t} (t-s)^{-\alpha} \left(v(s) - y_0\right) \, ds \right) \right]_{t=t_{ik}}$$
$$= \Gamma(1-\alpha) f(t_{ik}, v(t_{ik})), \quad i = 0, 1, ..., N-1, \ k = 1, 2, ..., \ell. \quad (12)$$

The natural way to proceed now is to consider the following type of Lagrange basis representation of v, whose restriction to σ_i, $i = 0, \ldots, N-1$, is (see [2]):

$$v\,(s) = \sum_{k=1}^{\ell} \mathscr{L}_{ik}(s) v(t_{ik}), \quad s \in \sigma_i, \quad (13)$$

where the Lagrange functions $\mathscr{L}_{ik} \in V^{\alpha}_{m}$ are defined by

$$\mathscr{L}_{ik}(t) = \sum_{p=1}^{\ell} \beta^{i}_{pk} t^{\nu_p}, \ i = 0, 1, ..., N-1, \ k = 1, ..., \ell, \quad (14)$$

and the coefficients $[\beta^{i}_{pk}]_{p=1,\ldots,\ell}$ may be determined by solving the $(\ell \times \ell)$ linear system of equations $\mathscr{L}_{ik}(t_{ij}) = \delta_{jk}, \ k, j = 1, \ldots, \ell$.

On the other hand, the approximate solution $v \in V^{\alpha}_{h,m}$ is such that $v(0) = y_0$, so that the number of unknowns is not less than the number of equations, we consider in the first subinterval, σ_0, the collocation parameters $c_0 = 0 < c_1 < ... < c_m \le 1$ and $v(s)$, for $s \in \sigma_0$, is given by

$$v(s) = \mathscr{L}_{01}(s) y_0 + \sum_{k=2}^{\ell} \mathscr{L}_{0k}(s) v(t_{0k}). \quad (15)$$

In this way, (11) and (12) becomes:

$$
\begin{aligned}
&y_0 \sum_{p=1}^{\ell} \gamma_{1p}^{1k} \beta_{p1}^{1} + \sum_{m=2}^{\ell} \sum_{p=1}^{\ell} \gamma_{1p}^{1k} \beta_{pm}^{1} v(t_{0m}) - y_0 t_{0k}^{-\alpha} \Gamma(1-\alpha) = f(t_{0k}, v(t_{0k})), \quad k = 2, ..., m, \\
&\sum_{j=0}^{i-1} \left(\sum_{m=1}^{\ell} \sum_{p=1}^{\ell} \gamma_{jp}^{ik} \beta_{pm}^{j} v(t_{jm}) \right) - y_0 t_{ik}^{-\alpha} + \sum_{m=1}^{\ell} \sum_{p=1}^{\ell} \gamma_{ip}^{ik} \beta_{pm}^{i} v(t_{im}) = \Gamma(1-\alpha) f(t_{ik}, v(t_{ik})), \\
&i = 2, ..., N-1, k = 1, 2, ..., m,
\end{aligned}
\tag{16}
$$

where the coefficients $\gamma_{j,m}^{ik}$ are defined through:

$$
\gamma_{jp}^{ik} = \begin{cases} \left[g(\nu_p, i, k, t_j) - g(\nu_p, i, k, t_{j+1}) \right] + \eta_p t_{ik}^{-\alpha+\nu_p} \left[h(\nu_p, t_j, i, k) - h(\nu_p, t_{j+1}, i, k) \right], & j \neq i, \\ g(\nu_p, i, k, t_i) + \eta_p t_{ik}^{-\alpha+\nu_p} \left[\dfrac{\Gamma(1-\alpha)\Gamma(\nu_p+1)}{\Gamma(2-\alpha-\nu_p)} - h(\nu_p, t_i, i, k) \right], & j = i, \end{cases}
$$

with

$$
\begin{aligned}
\eta_p &= (1 - \alpha + \nu_p), \\
g(\beta, i, k, t) &= t_{ik}^{-\alpha+\beta-1} t \left(1 - \frac{t}{t_{ik}}\right)^{-\alpha} \left(\frac{t}{t_{ik}}\right)^{\beta}, \\
h(\beta, t, i, k) &= B_{\frac{t}{t_{ik}}}(\beta + 1, 1 - \alpha),
\end{aligned}
$$

where $B_z(a, b)$ is the incomplete beta function.

Then, the solution of (1)–(2) restricted to the interval σ_i, $i = 0, 1, \ldots, N-1$, is given by (13) where the coefficients $v(t_{ik})$, $k = 1, 2, \ldots, \ell$, are the solution of the system above.

Obviously this is a linear system whenever $f(t, y(t)) = g(t) + by(t)$, for some real b and continuous function g.

3 Nonpolynomial Collocation Method Applied to the Time Fractional Diffusion Equation

In this section we apply the nonpolynomial collocation method to solve a time fractional diffusion equation, by using a combination of the method of lines and the proposed collocation method described on the previous section.

Let us consider the time fractional diffusion equation given by:

$$
\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} = D_\alpha \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad t > 0, \quad 0 \leq x \leq L, \tag{17}
$$

with initial condition:

$$u(x, 0) = g(x), \tag{18}$$

and boundary conditions:

$$u(0, t) = \varphi_0(t), \;\; u(L, t) = \varphi_L(t), \quad t > 0, \tag{19}$$

where D_α is the diffusion coefficient and α is the order of the fractional derivative, given in the Caputo sense, that satisfies $0 < \alpha < 1$.

We will use the method of lines for the numerical approximation of (17)–(19).

First, spatial derivatives are approximated using finite differences and then the resulting system of semi-discrete ordinary differential equations is solved by using the nonpolynomial collocation method described in the previous section.

We consider a uniform space mesh on the interval $[0, L]$, defined by the gridpoints $x_i = ih$, $i = 0, \ldots, n$, where $h = \frac{L}{n}$, and we approximate the space derivative by a second order finite difference:

$$\frac{\partial^2 u(x_i, t)}{\partial x^2} = \frac{u(x_{i+1}, t) - 2u(x_i, t) + u(x_{i-1}, t)}{h^2} + \mathcal{O}(h^2), \quad i = 1, \ldots, n-1. \tag{20}$$

We then obtain the following system of $n - 1$ ordinary fractional differential equations of order α:

$$\frac{\partial^\alpha y_i(t)}{\partial t^\alpha} = D_\alpha \frac{y_{i+1}(t) - 2y_i(t) + y_{i-1}(t)}{h^2} + f(x_i, t), \quad i = 1, \ldots, n-1, \tag{21}$$

where $y_i(t) \approx u(x_i, t)$.

Note that from the boundary conditions (19), we have $y_0(t) = \varphi_0(t)$, $y_n(t) = \varphi_L(t)$ and from the initial condition (18), we obtain:

$$y_i(0) = g(x_i), \;\; i = 1, \ldots, n-1, \tag{22}$$

and therefore, the solution of the $n - 1$ initial value problems (21)–(22) may be determined by using any initial value problem solver. The problem to solve may be outlined as follows.

For each $n \in \mathbf{N}$ and $t \geq 0$ we define

$$\mathbf{y}(t) = \left[y_0(t)\; y_1(t)\; y_2(t)\; \ldots\; y_{n-1}(t)\; y_n(t)\right] = \left[\varphi_0(t)\; y_1(t)\; y_2(t)\; \ldots\; y_{n-1}(t)\; \varphi_L(t)\right].$$

Thus, the system (21)–(22) can be rewritten as follows

$$\begin{cases} \dfrac{\partial^\alpha y_i(t)}{\partial t^\alpha} = G_i(t, \mathbf{y}(t)), \quad i = 1, \ldots, n-1, \\ y_i(0) = g(x_i), \;\; i = 1, \ldots, n-1, \end{cases} \tag{23}$$

where each function G_i is defined by

$$G_i(t, \mathbf{y}(t)) = D_\alpha \frac{y_{i+1}(t) - 2y_i(t) + y_{i-1}(t)}{h^2} + f(x_i, t),\ i = 1, \ldots, n-1,\ t > 0. \quad (24)$$

Hence, we end up with a system of $(n-1)$ fractional ordinary differential equations to solve. Each one of these systems is solved by using the nonpolynomial collocation method (16). As we will see, the numerical results suggest that we are able to obtain an optimal order of convergence in time.

4 Numerical Results

In this section we illustrate the performance of the proposed nonpolynomial collocation method (16) (NPCM). In order to do this we consider several examples.

Throughout this section, ε_N and $\widehat{\varepsilon}_N$ represent the errors at the collocation and discretisation points, respectively, and p the experimental order of convergence:

$$\varepsilon_N = \max_{i=0,1,\ldots,N-1} \max_{k=1,\ldots,\ell} |y(t_{ik}) - y_N(t_{ik})|, \quad (25)$$

$$\widehat{\varepsilon}_N = \max_{i=0,1,\ldots,N} |y(t_i) - y_N(t_i)|, \quad (26)$$

$$p = \log\left(\frac{\varepsilon_N}{\varepsilon_{2N}}\right) / \log(2), \quad (27)$$

where $y_N(t) \in V^{\alpha}_{h,m}$ and $h = \frac{T}{N}$.

On the other hand, the set of collocation parameters that we consider to define the collocation points at the subintervals σ_i, $i = 1, \ldots, N-1$ are given by:

- $c_1 = 0.25$, $c_2 = 0.308658$, $c_3 = 0.691342$, $c_\ell = 0.8$, if $\ell = 4$;
- $c_1 = 0.0380602$, $c_2 = 0.2$, $c_3 = 0.308658$, $c_4 = 0.691342$, $c_\ell = 0.96194$, if $\ell = 5$;
- $c_1 = 0.108658$, $c_2 = 0.308658$, $c_3 = 0.591342$, $c_4 = 0.69$, $c_5 = 0.791342$, $c_\ell = 0.96194$, if $\ell = 6$;
- $c_1 = 0.1$, $c_2 = 0.25$, $c_3 = 0.308658$, $c_4 = 0.45$, $c_5 = 0.591342$, $c_6 = 0.691342$, $c_7 = 0.8$, $c_\ell = 0.96194$, if $\ell = 8$;
- $c_1 = 0.1$, $c_2 = 0.15$, $c_3 = 0.25$, $c_4 = 0.308658$, $c_5 = 0.45$, $c_6 = 0.591342$, $c_7 = 0.691342$, $c_8 = 0.8$, $c_\ell = 0.96194$, if $\ell = 9$.

The collocation parameters c_i, $i = 2, \ldots, \ell$ on the first interval, σ_0, are given by the same values and $c_1 = 0$.

First we consider a linear initial value problem (IVP) whose solution is regular:

$$\begin{cases} D^{\alpha}\left(y(t)\right) = \frac{6(4-\alpha)}{\Gamma(1-\alpha)} t^{3-\alpha} - \frac{y(t)}{C_{\alpha}} + t^3, \quad t \in (0,1] \\ y(0) = 0, \end{cases} , \tag{28}$$

where $C_{\alpha} = \prod_{i=1}^{4} (\alpha - i)$ and the exact solution is given by $y(t) = C_{\alpha} t^3$.

In Table 1 we present the numerical results related with the approximate solution of (28) obtained by the application of the NPCM on the space $V_{h,2}^{\alpha}$ for several values of h and for $\alpha = \frac{1}{3}$, $\alpha = \frac{1}{2}$ and $\alpha = \frac{2}{3}$.

In Table 2 the maximum of the errors at the collocation and mesh points obtained by the application of the NPCM, on the space $V_{h,3}^{\alpha}$, for $\alpha = \frac{1}{2}$ and $\alpha = \frac{2}{3}$, are listed.

From Table 2 we can see that the maximum of the errors, using the NPCM on the space $V_{h,3}^{\alpha}$ applied to the Example (28), at the mesh points and collocation points, converges to zero with order 3. On the other hand, from Table 1 we observe that the the maximum of the errors, at the collocation parameters, using the NPCM on the space $V_{h,2}^{\alpha}$ converges to zero with convergence order approximately 3.

Table 1 Maximum of the errors and experimental order of convergence for the NPCM, applied to the IVP (28) on the spaces $V_{h,2}^{\alpha}$, $h = 1/N$, for several values of h and α

N	$\alpha = 1/3$		$\alpha = 1/2$		$\alpha = 2/3$	
	ε_N	p	ε_N	p	ε_N	p
5	$1.394 \cdot 10^{-2}$	–	$8.977 \cdot 10^{-3}$	–	$3.458 \cdot 10^{-3}$	–
10	$1.766 \cdot 10^{-3}$	2.98	$1.142 \cdot 10^{-3}$	2.97	$4.414 \cdot 10^{-4}$	2.97
20	$2.231 \cdot 10^{-4}$	2.98	$1.445 \cdot 10^{-4}$	2.98	$5.591 \cdot 10^{-5}$	2.98
40	$2.813 \cdot 10^{-5}$	2.99	$1.823 \cdot 10^{-5}$	2.99	$7.048 \cdot 10^{-6}$	2.99
80	$3.540 \cdot 10^{-6}$	2.99	$2.293 \cdot 10^{-6}$	2.99	$8.857 \cdot 10^{-7}$	2.99
160	$4.450 \cdot 10^{-7}$	2.99	$2.879 \cdot 10^{-7}$	2.99	$1.111 \cdot 10^{-7}$	3.00

Table 2 Maximum of the errors and experimental orders of convergence for the NPCM, applied to the IVP (28) on the space $V_{h,3}^{\alpha}$, $h = 1/N$, for several values of h

N	$\alpha = 1/2$				$\alpha = 2/3$			
	$\widehat{\varepsilon}_N$	p	ε_N	p	$\widehat{\varepsilon}_N$	p	ε_N	p
5	$4.120 \cdot 10^{-4}$	–	$8.503 \cdot 10^{-4}$	–	$3.467 \cdot 10^{-5}$	–	$6.143 \cdot 10^{-5}$	–
10	$5.336 \cdot 10^{-5}$	2.94	$1.083 \cdot 10^{-4}$	2.97	$4.587 \cdot 10^{-6}$	2.92	$7.863 \cdot 10^{-6}$	2.97
20	$6.840 \cdot 10^{-6}$	2.96	$1.373 \cdot 10^{-5}$	2.98	$5.943 \cdot 10^{-7}$	2.95	$9.977 \cdot 10^{-7}$	2.98
40	$8.704 \cdot 10^{-7}$	2.97	$1.732 \cdot 10^{-6}$	2.99	$7.600 \cdot 10^{-8}$	2.97	$1.259 \cdot 10^{-7}$	2.99
80	$1.102 \cdot 10^{-7}$	2.98	$2.180 \cdot 10^{-7}$	2.99	$9.638 \cdot 10^{-9}$	2.98	$1.583 \cdot 10^{-8}$	2.99
160	$1.390 \cdot 10^{-8}$	2.99	$2.739 \cdot 10^{-8}$	2.99	$1.216 \cdot 10^{-5}$	2.99	$1.986 \cdot 10^{-9}$	2.99

Now we consider a class of linear initial value problems whose analytical solution is not smooth:

$$\begin{cases} D^{\alpha}\,(y)\,(t) = \frac{105\left(\frac{9}{2}-\alpha\right)\sqrt{\pi}}{16\Gamma(11/2-\alpha)} t^{\frac{7}{2}-\alpha} + 3t^{\frac{7}{2}} - 3y \quad t \in (0, T],\ \alpha \in (0, 1), \\ y(0) = 0, \end{cases} . \tag{29}$$

For each $\alpha \in (0, 1)$ the exact solution of the previous initial value problem is $y(t) = t^{7/2}$.

In Tables 3 and 4 we present the numerical results obtained by the application of the NPCM on the spaces $V^{\alpha}_{h,m}$, $m = 2, 3$, for several values of h. The results of Tables 3 and 4 suggest that the maximum of the errors at the collocation points, using the NPCM on the spaces $V^{\alpha}_{h,m}$, $m = 2, 3$, converges to zero with order 3 in both cases.

Note that the solutions of the examples (28) and (29) are such that $y \in V^{\alpha}_{4}(I)$. In Fig. 1 we illustrate this fact with the absolute error related with the approximate solution obtained with the NPCM on the space $V^{1/2}_{1/40,4}$, for Examples (28) and (29).

Table 3 Maximum of the errors and experimental order of convergence for the NPCM, applied to the IVP (28) on the spaces $V^{\alpha}_{h,2}$, $h = 1/N$, for several values of h and α

N	$\alpha = 1/3$		$\alpha = 1/2$		$\alpha = 2/3$	
	ε_N	p	ε_N	p	ε_N	p
5	$3.553 \cdot 10^{-4}$	–	$3.693 \cdot 10^{-4}$	–	$4.985 \cdot 10^{-4}$	–
10	$3.852 \cdot 10^{-5}$	3.21	$4.078 \cdot 10^{-5}$	3.18	$5.392 \cdot 10^{-5}$	3.21
20	$4.102 \cdot 10^{-6}$	3.23	$4.316 \cdot 10^{-6}$	3.24	$5.440 \cdot 10^{-6}$	3.31
40	$4.283 \cdot 10^{-7}$	3.26	$4.390 \cdot 10^{-7}$	3.30	$5.4236 \cdot 10^{-7}$	3.38
80	$4.382 \cdot 10^{-8}$	3.29	$4.318 \cdot 10^{-8}$	3.35	$4.887 \cdot 10^{-8}$	3.42
160	$4.395 \cdot 10^{-9}$	3.32	$4.132 \cdot 10^{-9}$	3.39	$4.472 \cdot 10^{-9}$	3.45

Table 4 Maximum of the errors and experimental orders of convergence for the NPCM, applied to the IVP (28) on the space $V^{\alpha}_{h,3}$, $h = 1/N$, for several values of h

N	$\alpha = 1/3$		$\alpha = 1/2$		$\alpha = 2/3$	
	ε_N	p	ε_N	p	ε_N	p
5	$7.850 \cdot 10^{-6}$	–	$9.691 \cdot 10^{-5}$	–	$5.604 \cdot 10^{-5}$	–
10	$8.505 \cdot 10^{-7}$	3.21	$1.115 \cdot 10^{-5}$	3.12	$6.236 \cdot 10^{-6}$	3.17
20	$9.013 \cdot 10^{8}$	3.24	$1.214 \cdot 10^{-6}$	3.20	$6.401 \cdot 10^{-7}$	3.28
40	$9.336 \cdot 10^{-9}$	3.27	$1.259 \cdot 10^{-7}$	3.27	$6.228 \cdot 10^{-8}$	3.36
80	$9.457 \cdot 10^{-10}$	3.30	$1.254 \cdot 10^{-8}$	3.33	$5.852 \cdot 10^{-9}$	3.41
160	$9.383 \cdot 10^{-11}$	3.33	$1.210 \cdot 10^{-9}$	3.37	$5.377 \cdot 10^{-10}$	3.44

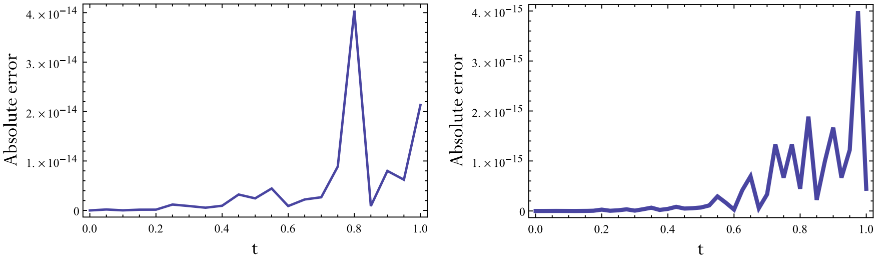


Fig. 1 Absolute errors related with the approximate solutions obtained with the NPCM on the space $V^{1/2}_{1/40,4}$. Left: Example (28). Right: Example (29).

Table 5 Maximum of the errors and experimental order of convergence for the NPCM, applied to the IVP (30) on the spaces $V^{\alpha}_{h,2}$, $h = 1/N$, for several values of h

N	$\alpha = 1/3$		$\alpha = 1/2$		$\alpha = 2/3$	
	ε_N	p	ε_N	p	ε_N	p
5	$1.576 \cdot 10^{-3}$	–	$3.026 \cdot 10^{-3}$	–	$1.160 \cdot 10^{-3}$	–
10	$1.080 \cdot 10^{-4}$	3.87	$3.574 \cdot 10^{-4}$	3.08	$6.184 \cdot 10^{-5}$	4.23
20	$7.044 \cdot 10^{-6}$	3.94	$3.581 \cdot 10^{-5}$	3.32	$3.178 \cdot 10^{-6}$	4.28
40	$4.421 \cdot 10^{-7}$	3.99	$3.328 \cdot 10^{-6}$	3.43	$1.605 \cdot 10^{-7}$	4.31
80	$2.691 \cdot 10^{-8}$	4.04	$2.992 \cdot 10^{-7}$	3.48	$8.034 \cdot 10^{-9}$	4.32
160	$1.600 \cdot 10^{-9}$	4.07	$2.653 \cdot 10^{-8}$	3.50	$4.003 \cdot 10^{-10}$	4.33

In what follows we consider a nonlinear example:

$$\begin{cases} D^{\alpha}(y)(t) = \frac{40320}{\Gamma(9-\alpha)} t^{8-\alpha} - 3\frac{\Gamma(5+\alpha/2)}{\Gamma(5-\alpha/2)} t^{4-\alpha/2} + \frac{9}{4}\Gamma(\alpha+1) \\ \qquad + \left(\frac{3}{2}t^{\alpha/2} - t^4\right)^3 - (y(t))^{3/2}, \quad t \in (0,1], \\ y(0) = 0. \end{cases} \tag{30}$$

The exact solution of this initial value problem is $y(t) = t^8 - 3t^{4+\alpha/2} + \frac{9}{4}t^{\alpha}$, meaning that the solution $y(t)$ can be written as $y(t) = u(t) + v(t)$ with $u(t) = \frac{9}{4}t^{\alpha} \in V^{\alpha}_{m}$ and $v(t) = t^8 - 3t^{4+\alpha/2} \in C^m([0,1])$, $m = 1, 2, 3, 4$.

For this example, the maximum of the errors and the experimental orders of convergence, related with the approximation obtained by the NPCM on the spaces $V^{\alpha}_{h,m}$, $m = 2, 3$, are presented, for different values of the stepsize h and α, in Tables 5 and 6.

Now we consider a multi-term initial value problem, a Bagley-Torvik fractional differential equation (see for example [1]). These equations arise, for example, in the modelling of the motion of a rigid plate immersed in a Newtonian fluid. Here we consider the Bagley-Torvik fractional differential equation given by:

$$\begin{cases} AD^2y(t) + BD^\nu y(t) + Cy(t) = g(t), & t \in (0, 1], \\ y(0) = 0, \ y'(0) = 0, \end{cases} \tag{31}$$

where $A = B = C = 1$, $\nu = 3/2$, $g(t) = 2 + \sqrt{t/\pi} + t^2$ and the analytical solution is known and given by $y(t) = t^2$.

First, we convert this problem into the equivalent linear system of equations

$$\begin{cases} D^{0.5}y_1(t) = y_2(t) \\ D^{0.5}y_2(t) = y_3(t) \\ D^{0.5}y_3(t) = y_4(t) \\ D^{0.5}y_4(t) = -y_1(t) - y_2(t) + g(t) \end{cases}, \tag{32}$$

together with the conditions $y_1(0) = 0$ and $y_3(0) = 0$.

On the other hand the derivative of non integer order, at $t = 0$, must be zero which implies $y_2(0) = 0$ and $y_4(0) = 0$.

In our numerical experiments we have used the NPCM, described in Sect. 2 and applied it to a system of differential equations of fractional order $\alpha = 1/2$, on the space $V^\alpha_{h,\ 2}$, to approximate the solution of the system of fractional equations (32). The numerical results presented in Table 7 suggest that the maximum of the errors at the collocation and mesh points converges to zero with the same order $p = 2$.

Table 6 Maximum of the errors and experimental orders of convergence for the NPCM, applied to the IVP (30) on the space $V^\alpha_{h,3}$, $h = 1/N$, for several values of h

N	$\alpha = 1/2$		$\alpha = 2/3$	
	ε_N	p	ε_N	p
5	$6.445 \cdot 10^{-4}$	–	$7.280 \cdot 10^{-4}$	–
10	$4.035 \cdot 10^{-5}$	4.00	$4.080 \cdot 10^{-5}$	4.15
20	$2.349 \cdot 10^{-6}$	4.10	$2.107 \cdot 10^{-6}$	4.28
40	$1.313 \cdot 10^{-7}$	4.16	$1.065 \cdot 10^{-7}$	4.31
80	$7.163 \cdot 10^{-9}$	4.20	$5.330 \cdot 10^{-9}$	4.32
160	$3.849 \cdot 10^{-10}$	4.22	$2.665 \cdot 10^{-10}$	4.32

Table 7 Maximum of the errors at the mesh and collocation points and estimates of the convergence order, p, using the NPCM on the space $V^\alpha_{1/N,2}$ to solve (32)

N	$\widehat{\varepsilon}_N$	p	ε_N	p
5	$1.455 \cdot 10^{-3}$	–	$5.088 \cdot 10^{-4}$	–
10	$3.649 \cdot 10^{-4}$	2.00	$1.326 \cdot 10^{-4}$	1.94
20	$9.129 \cdot 10^{-5}$	2.00	$3.350 \cdot 10^{-5}$	1.98
40	$2.283 \cdot 10^{-5}$	2.00	$8.400 \cdot 10^{-6}$	2.00
80	$5.707 \cdot 10^{-6}$	2.00	$2.102 \cdot 10^{-6}$	2.00

Table 8 Nonpolynomial collocation method on the spaces $V_{2,\tau}^{\alpha}$ and $V_{3,\tau}^{\alpha}$, $\alpha = 1/2$, for example (33) for several values of τ and h. The maximum of the absolute errors at the mesh points and the experimental orders of convergence p and q related with the stepsizes τ and h, respectively

$V_{2,\tau}^{\alpha}$			$V_{3,\tau}^{\alpha}$				
$h = \tau$	$\varepsilon_{h,\tau}$	$p = q$	h	τ	$\varepsilon_{h,\tau}$	p	p
1/4	$1.132 \cdot 10^{-2}$	–	1/4	1/3	$9.536 \cdot 10^{-3}$	–	–
1/8	$4.256 \cdot 10^{-3}$	1.41	1/8	1/4	$3.267 \cdot 10^{-3}$	1.56	2.32
1/16	$1.223 \cdot 10^{-3}$	1.80	1/16	1/7	$1.0495 \cdot 10^{-3}$	1.67	2.46
1/32	$3.263 \cdot 10^{-4}$	1.91	1/32	1/11	$2.922 \cdot 10^{-4}$	1.85	2.77
1/64	$8.418 \cdot 10^{-5}$	1.95	1/64	1/16	$7.709 \cdot 10^{-5}$	1.92	2.89

Finally the last example that we consider is an initial-boundary value problem for the time-fractional diffusion equation:

$$\begin{cases} \frac{\partial^{\alpha} u(x,t)}{\partial t^{\alpha}} = \frac{\partial^2 u(x,t)}{\partial x^2} + \frac{3\Gamma(\alpha)}{4} x^4 (x-1)t - 4x^2(5x-3)t^{1+\alpha}, \quad t > 0, \quad 0 \le x \le 1, \\ u(x,0) = 0, \\ u(0,t) = u(1,t) = 0. \end{cases} \tag{33}$$

whose analytical solution is $u(x,t) = x^4(x-1)t^{1+\alpha}$ and $\alpha = 1/2$.

The numerical error is measured by determining the maximum error at the mesh points (x_i, t_{jp}):

$$\varepsilon_{h,\tau} = \max_{\substack{i=1,\dots,n,\, j=1,\dots,N \\ p=1,\dots,\ell}} \left| u(x_i, t_{jp}) - y_i(t_{jp}) \right|, N = \frac{1}{\tau}, \; n = \frac{L}{h} \tag{34}$$

where y_i is the numerical solution obtained for the i-th spatial function and $u(x_i, t_{jp})$ is the exact solution evaluated at points (x_i, t_{jp}).

For each case, the experimental time and space rates of convergence were computed and denoted by p and q, respectively.

The numerical results obtained for the time-fractional diffusion equation (33) on the spaces $V_{m,\tau}^{\alpha}$, $m = 2, 3$, are presented in Table 8 and the numerical results suggest that $p \sim m$ (not dependent on the order of the fractional derivative) and $q \sim 2$.

5 Conclusions

In this work we have derived a piecewise nonpolynomial collocation method to approximate the solution of fractional differential equations that can deal with smooth and nonsmooth solutions of this type of problems. We also illustrate the accuracy and feasibility of the proposed method with several examples and the numerical results suggest that the proposed method is convergent and the order of convergence depends on the nonpolynomial space V_m^{α} that we choose to approximate the solution of the problem. The analysis of the convergence of the described numerical method will appear in a forthcoming paper. On the other hand, the numerical technique can also be used to solve a time fractional diffusion equation, by using a combination of the method of lines and the proposed piecewise nonpolynomial collocation method. For that problems the numerical results suggest that we obtain an optimal order of convergence in time.

Acknowledgements The authors acknowledge the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through projects MULTI/04621/2019 and UIDB/00297/2020 (Centro de Matemática e Aplicações), respectively.

References

1. Bagley, R.L., Torvik, P.J.: On the appearance of the fractional derivative in the behavior of real materials. ASME Trans. J. Appl. Mech. **51**, 294–298 (1984)
2. Cao, Y., Herdman, T., Xu, Y.: A hybrid collocation method for Volterra integral equations with weakly singular kernels. SIAM J. Numer. Anal. **41**(1), 364–381 (2003)
3. Caputo, M.: Elasticity e Dissipazione. Zanichelli, Bologna (1969)
4. Diethelm, K.: The Analysis of Fractional Differential Equations: An Application-Oriented Exposition Using Differential Operators of Caputo Type. Springer, Heidelberg (2004)
5. Li, C., Zeng, F.: Numerical Methods for Fractional Calculus. Numerical Analysis and Scientific Computing. Chapman & Hall/CRP Press, Boca Raton (2015)
6. Ford, N.J., Morgado, M.L., Rebelo, M.: Nonpolynomial collocation approximation of solutions to fractional differential equations. Fract. Calc. Appl. Anal. **16**(4), 874–891 (2013)
7. Samko, S.G., Kilbas, A.A., Marichev, O.I.: Fractional Integrals and Derivatives: Theory and Applications. Gordon and Breach, Yverdon (1993)
8. Kopteva, N.: Error analysis of the L1 method on graded and uniform meshes for a fractional-derivative problem in two and three dimensions. Math. Comput. (2017). https://doi.org/10.1090/mcom/3410

On Conditions on the Potential in a Sturm– Liouville Problem and an Upper Estimate of its First Eigenvalue

S. Ezhak and M. Telnova

Abstract We consider a Sturm–Liouville problem with Dirichlet boundary conditions and a weighted integral condition on the potential which may have singularities of different orders at the end-points of the interval $(0, 1)$. In this article we give one extra integral condition which is required for existence of the first eigenvalue of this problem. We find the values of parameters of the weighted integral condition, for which the first eigenvalue exists. We use the variational method for finding the first eigenvalue. Showing that the first eigenvalue is not greater than π^2, we prove that for $0 < \gamma < 1, \alpha, \beta > 2\gamma - 1$, the upper estimate for the first eigenvalue is strictly less than π^2.

1 Introduction

Consider the Sturm–Liouville problem

$$y'' + Q(x)y + \lambda y = 0, \quad x \in (0, 1), \tag{1}$$

$$y(0) = y(1) = 0, \tag{2}$$

where Q belongs to the set $T_{\alpha,\beta,\gamma}$ of all measurable locally integrable functions on $(0, 1)$ with non–negative values such that the integral condition

$$\int_0^1 x^{\alpha}(1-x)^{\beta} Q^{\gamma}(x)dx = 1, \quad \gamma \neq 0, \tag{3}$$

$$\int_0^1 x(1-x)Q(x)dx < \infty \tag{4}$$

hold.

S. Ezhak · M. Telnova (✉)
Plekhanov Russian University of Economics, Stremyanny lane 36, Moscow 117997, Russia
e-mail: mytelnova@yandex.ru

S. Ezhak
e-mail: svetlana.ezhak@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_37

A function y is a *solution to problem* (1), (2) if it is absolutely continuous on the segment $[0, 1]$, satisfies (2), its derivative y' is absolutely continuous on any segment $[\rho, 1-\rho]$, where $0 < \rho < \frac{1}{2}$, and equality (1) holds almost everywhere in the interval $(0, 1)$.

This work is continuation of studies initiated by Y. V. Egorov and V. A. Kondratiev in [1] the eigenvalue estimates of Sturm–Liouville problems, in particular, of the problem for the equation $y'' + \lambda Q(x)y = 0$ with Dirichlet boundary conditions and a non-negative summable on $[0, 1]$ potential Q satisfying the condition $\|Q\|_{L_\gamma(0,1)} = 1$, $\gamma \neq 0$. The problem for the equation $y'' - Q(x)y + \lambda y = 0$ with Dirichlet boundary conditions and a summable on $(0, \pi)$ potential Q: $\|Q\|_{L_\gamma(0,\pi)} \leq t$, for $\gamma \geq 1$, $t \geq 1$ was considered in [2]. We study a problem of that kind provided the integral conditions contain weight functions.

In Theorem 1 of this work we prove that if condition (4) does not hold, then for any $0 \leq p \leq +\infty$ there is no solution y to Eq. (1) with properties $y(0) = 0,\ y'(0) = p$.

From the results of [3] (Chapter 1, §2, Theorem 3) it follows that $T_{\alpha,\beta,\gamma}$ is empty provided $\gamma < 0$, $\alpha \leq 2\gamma - 1$ or $\beta \leq 2\gamma - 1$, for other values $\alpha, \beta, \gamma, \gamma \neq 0$, the set $T_{\alpha,\beta,\gamma}$ is not empty. Thus, for $\gamma < 0$, $\alpha \leq 2\gamma - 1$ or $\beta \leq 2\gamma - 1$, there is no function Q satisfying (3) and (4) taken together and, as a consequence, the first eigenvalue of problem (1), (2) does not exist.

Consider the functional

$$R[Q, y] = \frac{\int_0^1 y'^2 dx - \int_0^1 Q(x)y^2 dx}{\int_0^1 y^2 dx}.$$

Condition (4) is sufficient for boundedness of $R[Q, y]$ from below. In Theorem 2 of this work we prove that for any $Q \in T_{\alpha,\beta,\gamma}$,

$$\lambda_1(Q) = \inf_{y \in H_0^1(0,1)\setminus\{0\}} R[Q, y].$$

For any $\alpha, \beta, \gamma, \gamma \neq 0$, for any $Q \in T_{\alpha,\beta,\gamma}$, we have

$$M_{\alpha,\beta,\gamma} = \sup_{Q \in T_{\alpha,\beta,\gamma}} \inf_{y \in H_0^1(0,1)\setminus\{0\}} R[Q, y] \leq \inf_{y \in H_0^1(0,1)\setminus\{0\}} \frac{\int_0^1 y'^2 dx}{\int_0^1 y^2\, dx} = \pi^2.$$

It was proved [4] that if $\gamma > 1$, $-\infty < \alpha, \beta < +\infty$ or $0 < \gamma \leq 1$, $\alpha \leq 2\gamma - 1$, $-\infty < \beta < +\infty$ ($\beta \leq 2\gamma - 1$, $-\infty < \alpha < +\infty$), then $M_{\alpha,\beta,\gamma} = \pi^2$; if $\gamma < 0$, $\alpha, \beta > 2\gamma - 1$, then $M_{\alpha,\beta,\gamma} < \pi^2$.

In this article we prove that if $0 < \gamma < 1$, $\alpha, \beta > 2\gamma - 1$, then $M_{\alpha,\beta,\gamma} < \pi^2$.

2 Results

Theorem 1. *For any real number λ let us consider Eq. (1), where Q is a measurable locally integrable function on* $(0, 1)$ *with non–negative values such that for any* $0 < \delta < 1$

$$\int_0^\delta xQ(x)dx = +\infty.$$

Then there is no solution y to Eq. (1) such that $y(0) = 0,\ y'(0) = p$ for any $0 \le p \le +\infty$.

In order to prove this theorem we use Remark 2.1 to Theorem 2.2 ([3], Chapter 1, §1): Let y be a solution to the equation $y'' + Q(x)y = 0$, where the function Q is measurable non–negative summable on any segment $[\varepsilon, 1 - \varepsilon]$, $0 < \varepsilon < \frac{1}{2}$. Let y be defined on $[0, \delta]$, $0 < \delta \le 1$, and $y(0) = 0$, $y'(0) = p$, $p > 0$. Then we have

$$\int_0^1 xQ(x)dx < +\infty.$$

Proof. Suppose that there is a solution y to Eq. (1) such that $y(0) = 0,\ y'(0) = p$ for some $p > 0$. Then in some right δ–semineighborhood of 0 we have $y(x) > \frac{p}{2}x$. Since

$$y'(\delta) = p + \int_0^\delta -(Q(x) + \lambda)ydx,$$

then we obtain that the integral

$$\int_0^\delta -(Q(x) + \lambda)ydx = y'(\delta) - p$$

is finite. The integral $\int_0^\delta \lambda ydx$ is also finite. Consequently, the integral $\int_0^\delta Q(x)ydx$ is finite, too, but on the contrary

$$\int_0^\delta Q(x)ydx > \frac{p}{2}\int_0^\delta Q(x)xdx = +\infty.$$

Consequently, if the first eigenfunction of the problem exists provided $\int_0^1 xQ(x)dx = +\infty$, then $y'(0) = +\infty$ or $y'(0) = 0$.

If $y'(0) = +\infty$, then there exsists a right δ–semineighborhood of 0, where $y'' \le 0$. Then in this neighborhood we have $Q(x) + \lambda \ge 0$ and taking into account Remark 2.1 we obtain again that

$$\int_0^\delta xQ(x)dx < +\infty.$$

Assume that $y'(0) = 0$. Since $y(0) = 0$ and y is non-negative, then in some right δ–semineighborhood of 0 the function y can be convex downward and $y'' \geq 0$. Then in this neighborhood we have $Q(x) + \lambda \leq 0$. If $0 \leq Q(x) \leq -\lambda$, then Q is bounded on $[0, \delta]$, and again we have

$$\int_0^\delta xQ(x)dx < +\infty.$$

Under the conditions $y(0) = 0$ and $y'(0) = 0$, in some right δ–semineighborhood of 0 the derivative of the second order y'' can change its sign, moreover, there is no interval $(0, \delta_1)$, $\delta_1 < \delta$, at all points of which $y''(x) > 0$.

Integrating equality (1) over $[0, \delta]$, we obtain the equation

$$-y'(\delta) - \int_0^\delta \lambda y dx = \int_0^\delta Q(x) y dx. \tag{5}$$

If equality (5) holds, then $-y'(\delta) - \int_0^\delta \lambda y dx$ is a positive number M. If y'' changes its sign and y' exists on $(0, 1)$, then there are points in δ–semineighborhood of 0, at which y has minimum. Let us numerate these minimum-points, taking the rightmost one, which is less than δ, as x_1. We obtain the sequence of minimum-points, which converges to 0 from the right. Let us join by segments the neighbor points of the graph of y which correspond to these minimum-points. On each segment Δ_i between neighbor minimum-points x_{i+1} and x_i, $i \geq 1$, the graph of the function lies above one of the lines joining points $(0, 0)$ and $(x_{i+1}, y(x_{i+1}))$ or points $(0, 0)$ and $(x_i, y(x_i))$. The equation of this line is $y = p_i x$, where $p_i = \min\{y(x_{i+1}), y(x_i)\}$.

In virtue of the Lebesgue integral property ([8], Theorem 8, p. 141) if $f(x)$ is a summable on the set E function, then for any $\varepsilon > 0$ there is a number $\delta > 0$ such that for any set $e \subset E$ the inequality $\mu(e) < \delta$ implies that $\left|\int_e f(x)dx\right| < \varepsilon$.

Since $\int_0^\delta Q(x)xdx = +\infty$, then there is a positive number $\varepsilon = \frac{M}{p_*}$, such that for any $\delta > 0$ there exists a set $e \subset E$, such that $\mu(e) < \delta$ and

$$\left|\int_e Q(x)xdx\right| \geq \varepsilon.$$

Thus, choosing a set e as a set belonging to one of the segments Δ_i and denoting by p_* the corresponding p_i, we obtain

$$\int_0^\delta Q(x)ydx > \sum_{i=1}^\infty \int_{\Delta_i} Q(x)p_i xdx > p_* \int_e Q(x)xdx \geq M,$$

and equality (5) does not hold.

Theorem 1 is proved.

Theorem 2. *For any* $\gamma < 0$, $\alpha, \beta > 2\gamma - 1$ *and* $\gamma > 0$, $-\infty < \alpha, \beta < +\infty$, *for any function* $Q \in T_{\alpha,\beta,\gamma}$,

$$\lambda_1(Q) = \inf_{y \in H_0^1(0,1)\setminus\{0\}} R[Q, y].$$

Proof. By the Hölder inequality, for any $y \in H_0^1(0, 1)$ and $x \in (0, 1)$, we have

$$y^2(x) = \left(\int_0^x y'(t)\,dt\right)^2 \le x \int_0^x y'^2(t)\,dt,$$

$$y^2(x) = \left(-\int_x^1 y'(t)\,dt\right)^2 \le (1 - x) \int_x^1 y'^2(t)\,dt.$$

Then

$$\frac{y^2}{x(1-x)} = \frac{y^2}{x} + \frac{y^2}{1-x} \le \int_0^x y'^2(t)\,dt + \int_x^1 y'^2(t)\,dt = \int_0^1 y'^2(t)\,dt,$$

$$\int_0^1 Q(x)y^2dx \le \left(\int_0^1 y'^2dx\right) \int_0^1 x(1-x)Q(x)dx$$

and

$$R[Q, y] = \frac{\int_0^1 y'^2dx - \int_0^1 Q(x)y^2dx}{\int_0^1 y^2dx} \ge \frac{\int_0^1 y'^2dx \left(1 - \int_0^1 x(1-x)Q(x)dx\right)}{\int_0^1 y^2dx}.$$

Since $R[Q, y] = R[Q, |y|]$, therefore, we can assume that the function y is nonnegative. If any function $y \in H_0^1(0, 1)$ is convex downward on some subsegment of $[0, 1]$, we can construct the other function y_1, convex upward on $[0, 1]$ and, as a consequence, positive on $(0, 1)$ such that $R[Q, y_1] \le R[Q, y]$. To do it we join by segment any two points on the graph of y, the leftmost and the rightmost ones, between which the function is convex downward. Therefore, investigating the boundedness of the functional R from below, we can consider functions which are positive on $(0, 1)$ and convex upward.

Let

$$\Gamma_* = \left\{y \in H_0^1(0, 1) \mid \int_0^1 y^2\,dx = 1\right\},$$

$$I[Q,y]=\int\limits_0^1 y'^2\,dx-\int_0^1 Q(x)y^2dx.$$

Since for any $y\in H_0^1(0,1)$ the equality $R[Q,y]=R\left[Q,\frac{y}{\int_0^1 y^2dx}\right]$ holds, then

$$\inf_{y\in H_0^1(0,1)\setminus\{0\}} R[Q,y]=\inf_{y\in\Gamma_*} I[Q,y].$$

For any $y\in H_0^1(0,1)$, we have

$$\int_0^1 y^2dx\le\frac{1}{2}\int_0^1 y'^2dx.$$

Then for any $y\in\Gamma_*$,

$$\int_0^1 y'^2dx\ge 2.$$

If the function $y\in\Gamma_*$ is convex upward, then on $\left[0,\frac{1}{2}\right]$ we have

$$y(x)\ge 2y\left(\frac{1}{2}\right)\cdot x$$

and on $\left[\frac{1}{2},1\right]$ we have

$$y(x)\ge 2y\left(\frac{1}{2}\right)\cdot(1-x).$$

Consider the function

$$\widetilde{y}(x)=\begin{cases}2y\left(\frac{1}{2}\right)\cdot x, & x\in\left[0,\frac{1}{2}\right],\\ 2y\left(\frac{1}{2}\right)\cdot(1-x), & x\in\left[\frac{1}{2},1\right].\end{cases}$$

Since y is positive on $(0,1)$, we can write the following relations

$$\begin{gathered}\int_0^1 Q(x)y^2dx\le\sup_{[0,1]}\frac{y^2}{x(1-x)}\int_0^1 Q(x)x(1-x)dx\\ =\sup_{[0,1]}\frac{1}{\frac{x(1-x)}{y^2}}\int_0^1 Q(x)x(1-x)dx.\end{gathered}$$

Further,

$$\sup_{[0,1]}\frac{1}{\frac{x(1-x)}{y^2}}<\sup_{[0,1]}\frac{1}{\frac{x(1-x)}{\widetilde{y}^2}}=\frac{1}{\inf_{[0,1]}\frac{x(1-x)}{\widetilde{y}^2}}.$$

On $\left[0,\frac{1}{2}\right]$

$$\inf_{[0,\frac{1}{2}]}\frac{x(1-x)}{\widetilde{y}^2}=\inf_{[0,\frac{1}{2}]}\frac{x(1-x)}{4y^2\left(\frac{1}{2}\right)x^2}=\inf_{[0,\frac{1}{2}]}\frac{1-x}{4y^2\left(\frac{1}{2}\right)x}=\frac{\frac{1}{2}}{4y^2\left(\frac{1}{2}\right)\frac{1}{2}}=\frac{1}{4y^2\left(\frac{1}{2}\right)}.$$

On $\left[\frac{1}{2},1\right]$

$$\inf_{[\frac{1}{2},1]}\frac{x(1-x)}{\widetilde{y}^2}=\inf_{[\frac{1}{2},1]}\frac{x(1-x)}{4y^2\left(\frac{1}{2}\right)(1-x)^2}=\inf_{[\frac{1}{2},1]}\frac{x}{4y^2\left(\frac{1}{2}\right)(1-x)}=\frac{1}{4y^2\left(\frac{1}{2}\right)}.$$

Then

$$\int_0^1 Q(x)y^2dx \le 4y^2\left(\frac{1}{2}\right)\int_0^1 Q(x)x(1-x)dx.$$

The function y is convex upward, consequently, the area of the region bounded by the graph of y and the x-axis is greater than the area of the triangle with the vertices $(0,0)$, $(\frac{1}{2}, y\left(\frac{1}{2}\right))$, $(1,0)$. Therefore,

$$1=\left(\int_0^1 y^2dx\right)^{\frac{1}{2}}\ge\int_0^1 ydx>\frac{1}{2}y\left(\frac{1}{2}\right),$$

and

$$y\left(\frac{1}{2}\right)<2.$$

Thus,

$$4y^2\left(\frac{1}{2}\right)\int_0^1 Q(x)x(1-x)dx<16\int_0^1 Q(x)x(1-x)dx=Const$$

and

$$\begin{aligned}\textstyle\int_0^1 y'^2dx-\int_0^1 Q(x)y^2dx &\ge \textstyle\int_0^1 y'^2dx-4y^2\left(\frac{1}{2}\right)\int_0^1 Q(x)x(1-x)dx\\ &> \textstyle 2-16\int_0^1 Q(x)x(1-x)dx.\end{aligned}$$

If the integral $\int_0^1 x(1-x)Q(x)dx$ is finite, then $I[Q,y]$ is bounded from below in Γ_* and $R[Q,y]$ is bounded from below in $H_0^1(0,1)$. Thus, for $\gamma<0$, $\alpha,\beta>2\gamma-1$ and $\gamma>0$, $-\infty<\alpha,\beta<+\infty$, for any $Q\in T_{\alpha,\beta,\gamma}$, the functional $R[Q,y]$ is bounded from below in $H_0^1(0,1)$ and there is a finite

$$\inf_{y\in H_0^1(0,1)\backslash\{0\}} R[Q,y]=m.$$

Lemma 1. *For any $\gamma<0$, $\alpha,\beta>2\gamma-1$ and $\gamma>0$, $-\infty<\alpha,\beta<+\infty$, for any $Q\in T_{\alpha,\beta,\gamma}$, there exists a positive on $(0,1)$ function $u\in\Gamma_*$ such that*

$$R[Q,u] = \inf_{y\in\Gamma_*} I[Q,y].$$

Proof. Let $\{\widetilde{q}_k\}$ be a minimizing sequence of the functional $R[Q,y]$ in $H_0^1(0,1)$. Then $\{q_k\} = \{\frac{\widetilde{q}_k}{C_k^{1/2}}\}$, where $C_k = \int\limits_0^1 \widetilde{q}_k{}^2\,dx$, is a minimizing sequence of the functional $I[Q,y]$ in Γ_*, i.e. $I[Q,q_k] \to m$ as $k \to \infty$.

We can assume that q_k is non-negative. If any function q_k is convex downward on some subsegment of $[0,1]$, we can construct the other function y_k, convex upward on $[0,1]$ and positive on $(0,1)$ such that $R[Q,y_k] \le R[Q,q_k]$. Therefore, let us assume that every function of the minimizing sequence $\{y_k\}$ is positive on $(0,1)$ and convex upward.

Let us show that the sequence $\{y_k\}$ is bounded in $H_0^1(0,1)$. By one and the same reasons we can show that for any k,

$$\int_0^1 Q(x)y_k^2 \le 16\int_0^1 Q(x)x(1-x)dx.$$

For any sufficiently large k,

$$\int_0^1 y_k'^2dx - \int_0^1 Q(x)y_k^2dx < m+1$$

and

$$\int_0^1 y_k'^2dx < m+1+16\int_0^1 Q(x)x(1-x)dx = Const.$$

Since for $\gamma < 0, \alpha, \beta > 2\gamma - 1$ and $\gamma > 0, -\infty < \alpha, \beta < +\infty$, $\{y_k\}$ is bounded in $H_0^1(0,1)$, then it contains the subsequence $\{z_k\}$, which converges weakly in $H_0^1(0,1)$ to some function u, moreover, $\|u\|^2_{H_0^1(0,1)}$ is bounded by the same constant as $\|z_k\|^2_{H_0^1(0,1)}$. $H_0^1(0,1)$ is compactly embedded in $C[0,1]$, consequently, there exists a subsequence $\{s_k\}$ of $\{y_k\}$, which converges in $C[0,1]$. Since $C[0,1]$ is embedded in $L_2(0,1)$, the sequence $\{s_k\}$ converges in $L_2(0,1)$ to some function u and as $k \to \infty$,

$$\int\limits_0^1 s_k^2\,dx \longrightarrow \int\limits_0^1 u^2\,dx, \qquad \int\limits_0^1 u^2\,dx = 1. \tag{6}$$

Since $\{s_k\}$ is bounded in $H_0^1(0,1)$, the sequence $\{s_k'\}$ is bounded in $L_2(0,1)$. Then there exists a subsequence $\{w_k\}$ of $\{s_k\}$ such that the sequence $\{w_k'\}$ converges weakly to the function u' in $L_2(0,1)$. Then (see. [6, p. 217])

$$\|u'\|^2_{L_2(0,1)} \le \varliminf_{k\to\infty} \|w_k'\|^2_{L_2(0,1)} = A.$$

Thus,

$$\|u'\|^2_{L_2(0,1)} \le A. \tag{7}$$

Let $\{v_k\}$ be a subsequence of $\{w_k\}$ such that

$$\lim_{k\to\infty}\int_0^1 {v'_k}^2\,dx = \varliminf_{k\to\infty}\int_0^1 {w'_k}^2\,dx = A.$$

Since m is a limit of the sequence $\{I[Q, v_k]\}$, $m - A$ is a limit of the sequence $\left\{-\int_0^1 Q(x)v_k^2\,dx\right\}$. Then for any $\varepsilon > 0$ there exists a number K such that for any $k \ge K$,

$$\int_0^1 Q(x)v_k^2\,dx > A - m - \varepsilon. \tag{8}$$

Let us apply the Lebesgue theorem (see. [9, p. 5]). Since $\{v_k^2\}$ converges to u^2 in $L_1(0,1)$, there exists a subsequence $\{r_k\}$ of $\{v_k\}$ such that the sequence $\{Q(x)r_k^2\}$ converges to the function $Q(x)u^2$ as $k \to \infty$ almost everywhere in [0, 1]. We have shown that for any r_k,

$$\int_0^1 Q(x)r_k^2\,dx \le 16\int_0^1 Q(x)x(1-x)dx = Const.$$

Then

$$Q(x)u^2 \in L_1(0,1)$$

and as $k \to \infty$,

$$\int_0^1 Q(x)r_k^2\,dx \longrightarrow \int_0^1 Q(x)u^2\,dx.$$

If for any $k \ge K$, inequality (8) holds, then $\int_0^1 Q(x)u^2\,dx \ge A - m - \varepsilon$.

Since ε can be arbitrary small, we obtain $\int_0^1 Q(x)u^2\,dx \ge A - m$ and

$$-\int_0^1 Q(x)u^2\,dx \le m - A. \tag{9}$$

In virtue of (7) and (9), we get $I[Q, u] \le m$. Since $m = \inf\limits_{y \in \Gamma_*} I[Q, y]$, we have $I[Q, u] = m$. In virtue of (6), we obtain $u \in \Gamma_*$.

As a limit-function of $\{r_k\}$, u is non-negative on [0, 1]. Let us prove that u is convex upward and, consequently, is positive on (0, 1).

Assume that there exist points $x_1, x_2 \in [0, 1]$ and a number $0 < \mu < 1$ such that

$$u(\mu x_1 + (1 - \mu)x_2) < \mu u(x_1) + (1 - \mu)u(x_2).$$

Consider the function

$$\tilde{u}(x) = \begin{cases} u, & x \in [0, 1] \setminus [x_1, x_2], \\ u(x_1) + (x - x_1)\frac{u(x_2) - u(x_1)}{x_2 - x_1}, & x \in [x_1, x_2]. \end{cases}$$

We have $u(x_1) = \tilde{u}(x_1)$, $u(x_2) = \tilde{u}(x_2)$. Let $x_3 = \mu x_1 + (1 - \mu)x_2$. Then $u(x_3) < \tilde{u}(x_3)$. Put

$$\tilde{x}_1 = \sup_{u(x) \ge \tilde{u}(x),\, x_1 \le x < x_3} x, \qquad \tilde{x}_2 = \inf_{u(x) \ge \tilde{u}(x),\, x_3 < x \le x_2} x.$$

Then for any $x \in (\tilde{x}_1, \tilde{x}_2)$, we have $u(x) < \tilde{u}(x)$.

Put

$$\widehat{u}(x) = \begin{cases} u, & x \in [0, 1] \setminus [\tilde{x}_1, \tilde{x}_2], \\ u(\tilde{x}_1) + (x - \tilde{x}_1)\frac{u(\tilde{x}_2) - u(\tilde{x}_1)}{\tilde{x}_2 - \tilde{x}_1}, & x \in [\tilde{x}_1, \tilde{x}_2]. \end{cases}$$

Let us show that $[Q, \widehat{u}] < R[Q, u]$. Indeed,

$$\int_0^1 \widehat{u}^2 dx > \int_0^1 u^2 dx, \quad \int_0^1 Q(x)\widehat{u}^2 dx > \int_0^1 Q(x)u^2 dx,$$

$$\int_0^1 \widehat{u}'^2 dx \le \int_0^1 u'^2 dx.$$

The latter inequality follows from the equality

$$\int_0^1 \widehat{u}'^2 dx - \int_0^1 u'^2 dx = \int_{\tilde{x}_1}^{\tilde{x}_2} \widehat{u}'^2 dx - \int_{\tilde{x}_1}^{\tilde{x}_2} u'^2 dx$$

and the fact that, as $y(\tilde{x}_1) = u(\tilde{x}_1)$, $y(\tilde{x}_2) = u(\tilde{x}_2)$, the minimum of the functional $J[y] = \int_{\tilde{x}_1}^{\tilde{x}_2} y'^2 dx$ is attained at the function $y = C_1 x + C_2$, where C_1, C_2 are constants. Therefore, $R[Q, \widehat{u}] < R[Q, u]$. Since $\widehat{u} \in H_0^1(0, 1)$, we get the contradiction with the fact that $\inf_{H_0^1(0,1)\setminus\{0\}} R[Q, y] = R[Q, u]$. Consequently, u is a positive convex upward on (0, 1) function.

Lemma 1 is proved.

Lemma 2. *Let the function u satisfy the conditions of Lemma 1. Then u is a solution to the equation*

$$y'' + Q(x)y + my = 0,$$

where m is the minimal eigenvalue of problem (1), (2).

Proof. Let $z \in H_0^1(0,1)$. Consider a function of variable $t \in \mathbf{R}$

$$g(t) = \frac{\int_0^1 (u' + tz')^2\,dx - \int_0^1 Q(x)(u+tz)^2\,dx}{\int_0^1 (u+tz)^2 dx}.$$

Since $g(0) = \inf_{H_0^1(0,1)\setminus\{0\}} R[Q,y] = R[Q,u]$, then $g'(0) = 0$. If $u \in \Gamma_*$ and $I[Q,u] = m$, then for any $z \in H_0^1(0,1)$,

$$\int_0^1 u'z'dx - \int_0^1 Q(x)uz\,dx = m\int_0^1 u\,z\,dx.$$

Note that for $\gamma < 0, \alpha, \beta > 2\gamma - 1$ and $\gamma > 0, -\infty < \alpha, \beta < +\infty$, for any $z \in H_0^1(0,1)$, the integral $\int\limits_0^1 Q(x)uz\,dx$ absolutely converges because

$$\int\limits_0^1 Q(x)|uz|\,dx \le \left(\int_0^1 u'^2 dx\right)^{\frac{1}{2}} \left(\int_0^1 z'^2 dx\right)^{\frac{1}{2}} \left(\int\limits_0^1 x(1-x)Q(x)\,dx\right).$$

If $z \in C_0^\infty(0,1)$, then u' has a generalized derivative

$$u'' = -Q(x)u - mu.$$

According to Corollary 2.6.1. of Theorem 2.6.1 (see. [7, p. 41]), if $u, v \in L_p(a,b)$, $1 \le p \le \infty$, and the interval (a,b) is finite, $v(x)$ is the generalized derivative of the k—th order of $u(x)$, then $u(x)$ is continuously differentiable $k-1$ times on $[a,b]$ and almost everywhere on it has the classical derivative of the k—th order $u^{(k)}(x) = v(x)$. Moreover, the derivative $u^{(k-1)}(x)$ is absolutely continuous on $[a,b]$.

Since Q is a locally integrable on $(0,1)$ function, it is integrable on any segment $[\rho, 1-\rho]$, where $0 < \rho < \frac{1}{2}$. Then u is continuously differentiable on $[\rho, 1-\rho]$ and almost everywhere on it has the classical derivative of the second order

$$u'' = -Q(x)u - mu.$$

Moreover, u' is absolutely continuous on $[\rho, 1-\rho]$.

Thus, Eq. (1) holds almost everywhere in $(0,1)$. The boundary conditions hold because u belongs to $H_0^1(0,1)$. Consequently, u is a solution to problem (1), (2) with the eigenvalue $\lambda = m$.

For any solution z of problem (1), (2) let us multiply the left and the right parts of Eq. (1) by z_1, where

$$z_1(x) = \begin{cases} z, \; x \in (\rho, 1-\rho), \\ 0, \; x \in [0, \rho] \cup [1-\rho, 1], \end{cases}$$

and, integrating by parts over $[\rho, 1-\rho]$, we obtain:

$$\int_\rho^{1-\rho} z'^2\, dx - \int_\rho^{1-\rho} Q(x) z^2\, dx = \lambda \int_\rho^{1-\rho} z^2 dx.$$

Coming to the limit as $\rho \to 0$, we obtain the equality

$$\int_0^1 z'^2\, dx - \int_0^1 Q(x) z^2\, dx = \lambda \int_0^1 z^2 dx.$$

Here we use the facts that $z_1(\rho) = z_1(1-\rho) = 0$, $z'(\rho)$, $z'(1-\rho)$ are finite. Taking into account that $m = \inf\limits_{y \in H_0^1 \setminus \{0\}} R[Q, y]$, we obtain the inequality $\lambda \geq m$, which implies that m is the minimal eigenvalue of problem (1), (2).

Lemma 2 is proved.

Theorem 2 is proved.

Theorem 3. *If* $0 < \gamma < 1$, $\alpha, \beta > 2\gamma - 1$, *then* $M_{\alpha,\beta,\gamma} < \pi^2$.

Remark 1. For $0 < \gamma < 1/2$, the result $M_{0,0,\gamma} < \pi^2$ was obtained by A. Vladimirov [5]. The proof of Theorem 3 is based on [5], although the proofs of statements in the present work may differ from those supposed by the author in [5].

Proof. Let $0 < \gamma < 1$, $\alpha, \beta > 2\gamma - 1$ and $Q \in T_{\alpha,\beta,\gamma}$ be a function such that $\lambda_1(Q) > (\pi - \varepsilon)^2$, where $\varepsilon > 0$ is a sufficiently small number.

Consider the functions $\rho, \theta \in C^1[0,1]$ $y = \rho \cdot \sin\theta$, $y'/\sqrt{\lambda_1(Q)} = \rho \cdot \cos\theta$, and a measurable non–negative locally integrable on $(0, 1)$ function $\sigma = Q \cdot \sin^2\theta$.

Let us further write λ instead of $\lambda_1(Q)$. For the function y, we have

$$y'_x = \rho'_\theta \cdot \theta'_x \cdot \sin\theta + \rho \cdot \cos\theta \cdot \theta'_x = \theta'_x \left(\rho'_\theta \cdot \sin\theta + \rho \cdot \cos\theta\right) = \sqrt{\lambda} \cdot \rho \cdot \cos\theta,$$

$$y''_x = \sqrt{\lambda}\, (\rho'_\theta \cdot \theta'_x \cdot \cos\theta - \rho \cdot \sin\theta \cdot \theta'_x) = \sqrt{\lambda}\, \theta'_x\, (\rho'_\theta \cdot \cos\theta - \rho \cdot \sin\theta).$$

Having multiplied the first of these equalities by $\cos\theta$ and the second by $\sin\theta$ and subtracting, we obtain

$$-\frac{y''_x}{\sqrt{\lambda}} \sin\theta = \theta'_x\, \rho - \sqrt{\lambda}\, \rho\, \cos^2\theta.$$

On the other side,

$$y''_x = (-Q - \lambda)y = (-Q - \lambda)\rho \sin\theta = -\frac{\sigma\rho}{\sin\theta} - \lambda\rho\sin\theta$$

and

$$\theta'_x = \frac{\sigma + \lambda}{\sqrt{\lambda}}.$$

Since $\theta'(x) > 0$ on $(0, 1)$, the function θ is increasing on $[0, \pi]$. Under the assumptions of the theorem, we have

$$\int_0^1 \frac{\sigma\theta'}{\lambda + \sigma} dx = \int_0^1 \left(\theta' - \sqrt{\lambda}\right) dx = \pi - \sqrt{\lambda} < \varepsilon. \tag{10}$$

Estimate the integral

$$\int_0^1 x^\alpha (1-x)^\beta Q^\gamma(x)dx \leq A \int_0^{\frac{1}{2}} x^\alpha Q^\gamma(x)dx + B \int_{\frac{1}{2}}^1 (1-x)^\beta Q^\gamma(x)dx,$$

where

$$A = \begin{cases} 1, & \beta \geq 0, \\ \left(\frac{1}{2}\right)^\beta, & 2\gamma - 1 < \beta < 0, \end{cases} \qquad B = \begin{cases} 1, & \alpha \geq 0, \\ \left(\frac{1}{2}\right)^\alpha, & 2\gamma - 1 < \alpha < 0. \end{cases}$$

Consider the first of these integrals.

$$\begin{aligned} \textstyle\int_0^{\frac{1}{2}} x^\alpha Q^\gamma(x)dx &= \textstyle\int_0^{\frac{1}{2}} \sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha\, dx = \sqrt{\lambda} \int_0^{\frac{1}{2}} \frac{\sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha}{\lambda+\sigma}\theta' dx \\ &= \sqrt{\lambda}\left(\textstyle\int_{E_\varepsilon} \frac{\sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha}{\lambda+\sigma}\theta' dx + \int_{\overline{E_\varepsilon}} \frac{\sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha}{\lambda+\sigma}\theta' dx\right), \end{aligned}$$

where

$$E_\varepsilon = \{x \in \left[0, \frac{1}{2}\right] : \sigma(x) > \varepsilon^{\frac{\alpha - 2\gamma + 1}{1 - \gamma + \alpha}}\}.$$

Then

$$\begin{aligned} \textstyle\int_{E_\varepsilon} \theta' dx &= \textstyle\int_{E_\varepsilon} \frac{\sigma\theta'}{\lambda+\sigma} \cdot \frac{\lambda+\sigma}{\sigma} dx < \left(\frac{\lambda}{\varepsilon^{\frac{\alpha-2\gamma+1}{1-\gamma+\alpha}}} + 1\right) \int_{E_\varepsilon} \frac{\sigma\theta'}{\lambda+\sigma} dx \\ &< \left(\frac{\lambda}{\varepsilon^{\frac{\alpha-2\gamma+1}{1-\gamma+\alpha}}} + 1\right)\varepsilon \leq \pi^2 \cdot \varepsilon^{\frac{\gamma}{1-\gamma+\alpha}} + \varepsilon. \end{aligned}$$

Denote

$$\mu(\varepsilon) = \pi^2 \cdot \varepsilon^{\frac{\gamma}{1-\gamma+\alpha}} + \varepsilon < 2\pi^2 \cdot \varepsilon^{\frac{\gamma}{1-\gamma+\alpha}}.$$

For $\alpha > 2\gamma - 1$, the number $1 - \gamma + \alpha$ is positive. At the beginning of the proof we should have chosen ε such that $\mu(\varepsilon) \leq \frac{\pi}{2}$.

If $0 < \gamma \le \frac{1}{2}$, by the Cauchy inequality, knowing that $\lambda > 4$, we have

$$2\sqrt{\lambda}\sigma^{\gamma} = \frac{2}{\lambda^{\frac{1-2\gamma}{2}}}\,\lambda^{1-\gamma}\,\sigma^{\gamma} < \frac{2}{\lambda^{\frac{1-2\gamma}{2}}}\,((1-\gamma)\lambda + \gamma\sigma) < 2\lambda + 2^{\frac{1+2\gamma}{2}}\,\gamma\sigma \le 2\lambda + \sigma,$$

due to

$$\frac{2}{\lambda^{\frac{1-2\gamma}{2}}}\,(1-\gamma) \le 2, \qquad \frac{2}{\lambda^{\frac{1-2\gamma}{2}}}\,\gamma < 2^{\frac{1+2\gamma}{2}}\,\gamma \le 1.$$

If $\frac{1}{2} < \gamma < 1$, by the Cauchy inequality, knowing that $\lambda > 4$, we have

$$\begin{aligned} 2\sqrt{\lambda}\sigma^{\gamma} &= 2\lambda^{\gamma-\frac{1}{2}}\,\lambda^{1-\gamma}\,\sigma^{\gamma} < 2\lambda^{\gamma-\frac{1}{2}}\,((1-\gamma)\lambda + \gamma\sigma) \\ &< 2\pi\,\,((1-\gamma)\lambda + \gamma\sigma) < 2\pi\,\,(\lambda + \sigma)\,. \end{aligned}$$

In either case, for $0 < \gamma < 1$, we have

$$\sqrt{\lambda}\sigma^{\gamma} < \pi\,\,(\lambda + \sigma) \tag{11}$$

and

$$\sqrt{\lambda}\int_{E_\varepsilon} \frac{\sigma^{\gamma}\,\sin^{-2\gamma}\theta\,x^{\alpha}}{\lambda+\sigma}\theta' dx < \pi\int_{E_\varepsilon} \sin^{-2\gamma}\theta x^{\alpha}\theta' dx.$$

At first, let us consider the case $\alpha \ge 0$ and $\alpha > 2\gamma - 1$. The condition

$$\theta' = \frac{\lambda+\sigma}{\sqrt{\lambda}} > \sqrt{\lambda}$$

implies that

$$\theta(x) = \int_0^x \theta'(t)dt > \sqrt{\lambda}x.$$

Since on $\left[0, \frac{\pi}{2}\right]$ we have $\sin\theta \ge \frac{2}{\pi}\theta$, then

$$\sin^{-2\gamma}\theta \le \left(\frac{2}{\pi}\theta\right)^{-2\gamma} < \left(\frac{2}{\pi}\sqrt{\lambda}\right)^{-2\gamma} x^{-2\gamma}.$$

The inequality $\sqrt{\lambda} > 2$ implies that

$$x < \frac{\pi}{2}\frac{\sin\theta}{\sqrt{\lambda}} < \frac{\pi}{4}\sin\theta.$$

For $\alpha \ge 0$,

$$x^{\alpha} < \left(\frac{\pi}{4}\right)^{\alpha}\sin^{\alpha}\theta.$$

For $0 \le \alpha \le 2\gamma$ and $0 \le \mu(\theta) \le \frac{\pi}{2}$, since $\sin\theta \ge \frac{2}{\pi}\theta$, we have

$$\begin{aligned}\sqrt{\lambda}\int_{E_\varepsilon} \tfrac{\sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha}{\lambda+\sigma}\theta' dx &< \pi \int_{E_\varepsilon} \sin^{-2\gamma}\theta \cdot x^\alpha \theta' dx \\ \le \pi \int_0^{\mu(\varepsilon)} \left(\tfrac{\pi}{4}\right)^\alpha \left(\tfrac{2}{\pi}\right)^{\alpha-2\gamma} \theta^{\alpha-2\gamma} d\theta &< \pi^{2\gamma+1} 2^{-\alpha-2\gamma} \frac{\left(2\pi^2 \cdot \varepsilon^{\frac{\gamma}{1-\gamma+\alpha}}\right)^{\alpha-2\gamma+1}}{\alpha-2\gamma+1} \\ &= \tfrac{2^{1-4\gamma}\pi^{2\alpha-2\gamma+3}}{\alpha-2\gamma+1} \varepsilon^{\frac{\gamma(\alpha-2\gamma+1)}{1-\gamma+\alpha}}.\end{aligned}$$

Further, since $\frac{\sqrt{\lambda}}{\lambda+\sigma} < \frac{1}{\sqrt{\lambda}} < \frac{1}{2}$, we have

$$\begin{aligned}\sqrt{\lambda}\int_{\overline{E}_\varepsilon} \tfrac{\sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha}{\lambda+\sigma}\theta' dx &\le \tfrac{1}{2}\left(\tfrac{\pi}{4}\right)^\alpha \varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}} \int_0^\pi \sin^{\alpha-2\gamma}\theta \cdot d\theta \\ \le \tfrac{1}{2}\left(\tfrac{\pi}{4}\right)^\alpha \varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}} \int_0^\pi \left(\tfrac{2}{\pi}\right)^{\alpha-2\gamma} \theta^{\alpha-2\gamma} d\theta &= \pi^{\alpha+1} \tfrac{\varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}}}{\alpha-2\gamma+1} 2^{-\alpha-2\gamma-1}.\end{aligned}$$

Thus,

$$\int_0^{\frac{1}{2}} x^\alpha Q^\gamma(x)dx \le \frac{\varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}}}{\alpha-2\gamma+1}\left(2^{1-4\gamma}\pi^{2\alpha-2\gamma+3} + 2^{-\alpha-2\gamma-1}\pi^{\alpha+1}\right).$$

If $\alpha > 2\gamma$, then $\sin^{\alpha-2\gamma}\theta \le \theta^{\alpha-2\gamma}$ and

$$\int_0^{\frac{1}{2}} x^\alpha Q^\gamma(x)dx \le \frac{\varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}}}{\alpha-2\gamma+1}\left(2^{-\alpha-2\gamma+1}\pi^{3\alpha-4\gamma+3} + 2^{-2\alpha-1}\pi^{2\alpha-2\gamma+1}\right).$$

If $2\gamma - 1 < \alpha < 0$, consider the function

$$\theta(x) = \theta(0) + \int_0^x \theta'(t)dt = \int_0^x \theta'(t)dt.$$

Since θ is a continuously differentiable on $[0, \frac{1}{2}]$ function, the conditions of mean-value theorem hold and there exists a point $\xi \in [0, \frac{1}{2}]$ such that

$$\theta(x) = \theta(0) + \int_0^x \theta'(t)dt = \int_0^x \theta'(t)dt = \theta'(\xi)x \le \left(\max_{[0,\frac{1}{2}]}\theta'(x)\right)x.$$

Let $K = \max_{[0,\frac{1}{2}]}\theta'(x)$. Then on $[0, \frac{1}{2}]$ we have $\theta(x) \le Kx$. If $\alpha < 0$, then $x^\alpha \le K^{-\alpha}\theta^\alpha$. Since $\sin\theta \ge \frac{2}{\pi}\theta$, we obtain

$$\begin{aligned}\sqrt{\lambda}\int_{E_\varepsilon} \tfrac{\sigma^\gamma \cdot \sin^{-2\gamma}\theta \cdot x^\alpha}{\lambda+\sigma}\theta' dx &< \pi \int_{E_\varepsilon} \sin^{-2\gamma}\theta \cdot x^\alpha \theta' dx \\ \le \pi \int_0^{\mu(\varepsilon)} \left(\tfrac{2}{\pi}\right)^{-2\gamma} \theta^{\alpha-2\gamma} K^{-\alpha} d\theta &= \pi \left(\tfrac{2}{\pi}\right)^{-2\gamma} K^{-\alpha} \frac{\left(2\pi^2 \cdot \varepsilon^{\frac{\gamma}{1-\gamma+\alpha}}\right)^{\alpha-2\gamma+1}}{\alpha-2\gamma+1} \\ &= \tfrac{2^{\alpha-4\gamma+1}\pi^{2\alpha-2\gamma+3}}{\alpha-2\gamma+1} K^{-\alpha} \varepsilon^{\frac{\gamma(\alpha-2\gamma+1)}{1-\gamma+\alpha}},\end{aligned}$$

$$\sqrt{\lambda}\int_{\overline{E}_\varepsilon} \tfrac{\sigma^\gamma\cdot\sin^{-2\gamma}\theta\cdot x^\alpha}{\lambda+\sigma}\theta' dx \le \tfrac{1}{2}\varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}} \int_0^\pi K^{-\alpha}\theta^{\alpha-2\gamma}d\theta$$
$$= \tfrac{1}{2}K^{-\alpha}\varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}}\,\tfrac{\pi^{\alpha-2\gamma+1}}{\alpha-2\gamma+1}.$$

Thus,

$$\int_0^{\frac{1}{2}} x^\alpha Q^\gamma(x)dx \le \frac{\varepsilon^{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}}}{\alpha-2\gamma+1}K^{-\alpha}\pi^{\alpha-2\gamma+1}\left(2^{\alpha-4\gamma+1}\pi^{\alpha+2}+2^{-1}\right).$$

Similarly, changing variables $1-x=t$ or taking the fact that on $\left[\frac{\pi}{2},\pi\right]$

$$\sin\theta \ge \frac{2}{\pi}(\pi-\theta), \quad \theta(x) = \pi - \int_x^{\frac{1}{2}}\theta'(t)dt,$$

we obtain the similar result for $\int_{\frac{1}{2}}^1 (1-x)^\beta Q^\gamma(x)dx$.

For $\alpha, \beta > 2\gamma - 1$, the numbers $1-\gamma+\alpha$ and $1-\gamma+\beta$ are positive. Denote

$$M = \min\{\frac{(\alpha-2\gamma+1)\gamma}{1-\gamma+\alpha}, \frac{(\beta-2\gamma+1)\gamma}{1-\gamma+\beta}\}.$$

We have shown that there exists a constant $C > 0$ such that

$$\int_0^1 x^\alpha(1-x)^\beta Q^\gamma(x)dx \le C\varepsilon^M.$$

Since ε can be arbitrary small, we get the contradiction with condition (3).

References

1. Egorov, Yu., Kondratiev, V.: On Spectral Theory of Elliptic Operators. Operator Theory Advances and Applications, vol. 89, Birkhouser, Basel (1996)
2. Vinokurov, V.A., Sadovnichii, V.A.: On the range of variation of an eigenvalue when the potential is varied. Dokl. Math. **68**(2), 247–253 (2003). Translated from Doklady Akademii Nauk
3. Kuralbaeva, K.Z.: Some optimal estimates of eigenvalues of Sturm–Liouville problems. Thesis (1996)
4. Ezhak, S.S., Telnova, M.Y.: Estimates for the first eigenvalue of the Sturm–Liouville problem with potentials in weighted spaces. J. Math. Sci. **244**(2), 216–235 (2020)
5. Vladimirov, A.A.: On an a priori majorant for eigenvalues of Sturm–Liouville problems. arXiv:1602.05228
6. Liusternik, L.A., Sobolev, V.J.: Elements of Functional Analysis. Nauka, Moscow (1965). Translated from the Russian by Anthony E. Labarre Jr., Herbert Izbicki, H. Ward. Crowley Frederick Ungar Publishing Co., New York (1961)
7. Mikhlin, S.G.: Linear Partial Differential Equations. Vyss. Skola, Moscow (1977)
8. Natanson, I.P.: Theory of Functions of a Real Variable. Nauka, Moscow (1974)
9. Osmolovskii, V.G.: The Nonlinear Sturm–Liouville Problem. SPbGU, Saint-Petersburg (2003)

IFOHAM-A Generalization of the Picard-Lindelöff Iteration Method

Marta Sacramento, Cecília Almeida, and Miguel Moreira

Abstract IFOHAM (Iterative First order HAM) is an iterative technique based on the first order equation of the Homotopy Analysis Method (HAM). It can be shown that IFOHAM generalizes Picard-Lindeloff's iteration algorithm and can be used to solve nonlinear differential equations. In this work IFOHAM will be implemented in an symbolic computer environment and we will analyze and test its applicability to find series solutions of second order nonlinear differential equations with periodic solutions. In particular, we will show that the IFOHAM method is able to identify the fundamental frequencies as well as the amplitudes of such periodic solutions. Knowledge of these parameters is of particular importance in design and maintenance activities as it characterizes the oscillatory behavior of many real systems with nonlinear responses. The results of tests performed using the IFOHAM method will be compared with results available in the literature using the HAM as well as with results obtained using classical numerical techniques to solve differential equations.

1 Introduction

The HAM (Homotopy Analysis Method) was developed by Shijun Liao [4] and consists in an analytic approximation method [11] to solve nonlinear ordinary differential equations as well as partial differential equations. This technique is based on the concept of homotopy and transforms the original problem:

$$N\left[u\right]=0 \tag{1}$$

M. Sacramento · C. Almeida
Navy Research Center (CINAV), Naval Academy, 2810-001 Almada, Portugal
e-mail: marta.sofia.sacramento@marinha.pt

C. Almeida
e-mail: cecilia.branco.almeida@marinha.pt

M. Moreira (✉)
Navy Research Center (CINAV), Centre of Marine Technology and Ocean Engineering (CENTEC), Naval Academy, 2810-001 Almada, Portugal
e-mail: miguel.moreira@marinha.pt

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_38

into a family of problems characterized by the following linear differential equations:

$$\mathscr{L}\left[u_1(t)\right] = c_0\left[N\left[u_0(t)\right]\right] \tag{2}$$

$$\mathscr{L}\left[u_m(t) - u_{m-1}(t)\right] = c_0\mathscr{D}_{m-1}\left[N\left[\phi(t;q)\right]\right] \tag{3}$$

with $m \in \mathscr{N}$ and $m > 1$. The recursive resolution of (2) and (3) gives the successive terms of the requested solution:

$$u(t) = u_0(t) + \sum_{i=1}^{+\infty} u_i(t). \tag{4}$$

Note that $\mathscr{L}$ represents an appropriate linear operator; $u_0 = u_0(t)$ represents an initial solution guess of the original problem; c_0 represents an appropriate convergence control parameter; $\mathscr{D}_k$ represents the homotopic derivative operator of order k defined by:

$$\mathscr{D}_k = \frac{1}{k!}\frac{\partial^k}{\partial q^k}\bigg|_{q=0}. \tag{5}$$

Finally, $\phi(t;q)$ represents the homotopy Maclaurin series reading:

$$\phi(t;q) = u_0(t) + \sum_{n=1}^{+\infty} u_n(t)q^n, \quad q \in [0,1]. \tag{6}$$

In the works [6–8], the explanation of the applications details of HAM, as well as, numerous illustrative applications, both introductory and advanced, can be consulted.

It's important to mention that the HAM has the following features, that gives advantages over other asymptotic nonlinear problem solving techniques:

- Guarantee of convergence by adequately choosing c_0, the convergence control parameter;
- Flexibility on the choice of base functions and decide about the solution expression by adequately choosing $\mathscr{L}$ and the initial guess $u_0(t)$;
- Great generality of application ranging from solving weakly to strong nonlinear differential equations or even fractional differential equations;
- Ability to find important parameters, such as amplitude and frequency, of periodic solutions of nonlinear problems.

Of these facts, among others, the HAM has been widely applied by the scientific community in recent years in solving nonlinear problems.

2 The IFOHAM Method

The IFOHAM (Iterative First order HAM) is an asymptotic technique for solving nonlinear differential equations that generalizes the Picard-Lindelöff's method and is inspired by the HAM (Homotopy Analysis Method). Indeed, in addressing the first order Initial Value Problem (IVP)

$$\begin{cases} \frac{dx}{dt} = f(t, x) \\ x(t_0) = x_0^{(0)} \end{cases}, \tag{7}$$

based on HAM, one can find

$$u_1(t) = c_0 \mathscr{L}^{-1} \left[N \left[u_0(t) \right] \right] \tag{8}$$

from (2). Assuming the convergence of (4), one can conjecture that

$$u_0(t) + u_1(t) \tag{9}$$

leads to a better initial guess than the (postulated) original one, $u_0(t)$. We are thus led to the iterative algorithm we call IFOHAM (see [10], for a more detailed description of this method):

$$\begin{cases} u_0(t) = x_0^{(0)} \\ \mathscr{L}\left[u_{n+1}(t)\right] = c_0 \left[N\left[\sum_{k=0}^{n} u_k(t)\right]\right], \quad n \geq 0 \\ \text{with} \quad u_k(t_0) = 0, \quad \forall k \in \mathscr{N} \\ x_n = \sum_{k=0}^{n} u_k(t) \end{cases}. \tag{10}$$

It is demonstrated in [10] that, if

$$\mathscr{L}\left[h(t)\right] = \frac{dh}{dt}(t), \tag{11}$$

and

$$N\left[x\right] \equiv \frac{dx}{dt} - f(t, x), \tag{12}$$

then, the iterative algorithm IFOHAM (10), is equivalent to the following algorithm

$$\begin{cases} x_0(t) = x_0^{(0)} \\ x_{n+1}(t) = (1 + c_0)x_n(t) - c_0(x_0^{(0)} + \int_{t_0}^{t} f(\xi, x_n(\xi))d\xi), \quad n \geq 0 \end{cases} \tag{13}$$

which generalizes the Picard-Lindelöff's iterative method. Clearly, if $c_0 = -1$, this technique (13) coincides exactly with the aforementioned Picard-Lindelöff's iterative process, as we can immediately observe:

$$\begin{cases} x_0(t) = x_0^{(0)} \\ x_{n+1}(t) = x_0^{(0)} + \int_{t_0}^{t} f(\xi, x_n(\xi))d\xi, \quad n \geq 0 \end{cases} \tag{14}$$

It was shown in [10] that the parameter c_0 also influences the convergence speed of IFOHAM applied to the family of problems described by (12). So, the proper choice of c_0 allows to improve the performance of the IFOHAM method.

In [10] it is conjectured the applicability of IFOHAM in solving problems already addressed and solved by the method HAM, such as, second order problems. The applicability of IFOHAM in determining the fundamental frequency, as well as, amplitudes of periodic solutions of nonlinear problems is also conjectured. In order to address these issues we will focus here on the autonomous Duffing equation and on the van der Pol equation.

3 Autonomous Duffing Equation

The Duffing equation is typically a second order non-linear differential equation with a cubic polynomial stiffness term, as well as, a linear viscous type damping term [2]. This nonlinear equation was introduced by Georg Duffing in 1918 as a result of his work in forced vibrations in a system with a cubic softening nonlinearity. The autonomous Duffing equation can display an interesting self-oscillatory behavior.

Consider the autonomous Duffing equation studied in detail in [7] and described by the second-order nonlinear differential equation.

$$\frac{d^2x}{dt^2} + \lambda x + \varepsilon x^3 = 0, \tag{15}$$

with the initial conditions,

$$x(0) = x^* \tag{16}$$

and

$$\frac{dx}{dt}(0) = 0. \tag{17}$$

In Chapter 2 of [7] the qualitative study of this nonlinear problem is performed, through which the ranges of values of λ and ε are determined, so that the problem in question presents periodic solutions. The knowledge of the fundamental frequency ω, characteristic of such periodic solutions is particularly important. This is because not only enables the expression of the solution in terms of a well-suited trigonometric series, but it also has an enormous practical importance.

The determination of periodic solutions of different orders using HAM, the corresponding square means of the residual and the fundamental frequencies is performed by Liao in the aforementioned work for different parameter values λ, ε and $x(0) = x^*$.

Table 1 Parameters used in tests A, B and C

	λ	ε	x^*
A	$\frac{9}{4}$	1	1
B	0	1	1
C	4	-1	-1

In order to compare the results obtained with the IFOHAM and the HAM, we will use the values of the parameters λ, ε e $x(0) = x^*$ shown in Table 1.

3.1 Strategy for Determining the Frequency Using IFOHAM

As with using the HAM method, in approaching the problem (15) using IFOHAM, it becomes necessary to make the variable transformation $\tau = \omega t$ and define

$$x(t) = y(\tau), \tag{18}$$

to highlight, in the equivalent problem resulting from this change of variable, the fundamental frequency ω. Then,

$$\frac{d^2x}{dt^2} = \omega^2 \frac{d^2y}{d\tau^2}, \tag{19}$$

and the expression (15) can be rewritten (doing $\gamma = \omega^2$) in the form

$$\gamma \frac{d^2y}{d\tau^2} + \lambda y + \varepsilon y^3 = 0. \tag{20}$$

The corresponding initial conditions that must be satisfied by the function $y = y(\tau)$ will be

$$y(0) = x^* \tag{21}$$

and

$$\frac{dy}{d\tau}(0) = 0. \tag{22}$$

From $y = y(\tau)$, the sought solution $x = x(t)$, of the differential equation (15), will simply be

$$x = x(t) = y(\omega t). \tag{23}$$

3.2 Using the IFOHAM Algorithm

In order to apply the IFOHAM algorithm, define

$$N\left[\gamma, y\right] = \gamma \frac{d^2 y}{d\tau^2} + \lambda y + \varepsilon y^3 = 0, \tag{24}$$

and let's follow the procedures adopted by the HAM users regarding the establishment of the rule of construction of the expression's solution, choosing an appropriate linear operator $\mathscr{L}$ and choosing an initial guess $u_0 = u_0(\tau)$ of the sought solution. The motivation for such choices can be found in Chapter 2 of [7] and also, for example, in [5]. We will then have

$$\begin{cases} u_0(\tau) = x^* \cos\tau \\ \gamma_n \mathscr{L}\left[u_{n+1}(\tau)\right] = c_0\left[N\left[\gamma_n, \sum_{k=0}^{n} u_k(\tau)\right]\right], \quad n \geq 0 \\ \text{with} \quad u_k(\tau_0) = 0 \quad \text{and} \quad u'_k(\tau_0) = 0 \quad \forall k \in \mathscr{N} \\ y_n = \sum_{k=0}^{n} u_k(\tau) \end{cases}, \tag{25}$$

with

$$\mathscr{L}\left[f\right] = \frac{d^2 f}{d\tau^2} + f. \tag{26}$$

The procedure we will use to determine the successive γ_n terms, unknown in the process (25) will follow the approach taken and justified, for example in [5]. This procedure is based on the elimination, in each iterate of the so-called secular terms $\tau \cos\tau$ resulting from the resolution of the linear differential equation:

$$\gamma_n \mathscr{L}\left[u_{n+1}(\tau)\right] = c_0\left[N\left[\gamma_n, \sum_{k=0}^{n} u_k(\tau)\right]\right]. \tag{27}$$

That is, at each iteration, the choice of γ_n is made to ensure that the term u_{n+1}, to be determined, doesn't have secular plots of type $\tau \cos\tau$. The origin of this procedure, as referred to and cited in [5], dates back to works by Lindstedt, Bohlin, Poincaré and Gyldén.

If the process described is convergent it will lead to sequences

$$\gamma_n \to \gamma \quad \text{and} \quad y_{n+1} = \sum_{k=0}^{n+1} u_k(\tau) \to y \tag{28}$$

whose limits satisfy the Initial Value Problem (20) and (22).

To better understand the application of the IFOHAM to solve the problem under study, we will exemplify below the analytical procedures necessary to compute the first iterate. From (24) and (26), we can deduce:

$$\gamma_0(\frac{d^2u_1}{d\tau^2}+u_1)=c_0(\gamma_0\frac{d^2u_0}{d\tau^2}+\lambda u_0+\varepsilon u_0^3). \tag{29}$$

Since $u_0(\tau)=x^*\cos\tau$ and noting that $\cos^3\tau=\frac{1}{4}\cos 3\tau+\frac{3}{4}\cos\tau$ we will obtain the expression of the differential equation (30) to solve to find u_1 and γ_0:

$$\gamma_0(\frac{d^2u_1}{d\tau^2}+u_1)=c_0\left\{(-\gamma_0x^*+\lambda x^*+\frac{3\varepsilon(x^*)^3}{4})\cos\tau+\frac{\varepsilon(x^*)^3}{4}\cos 3\tau\right\}. \tag{30}$$

In order to avoid the emergence of the secular terms, $\tau\cos\tau$, in the general solution of (30), it will be sufficient to ensure that

$$-\gamma_0x^*+\lambda x^*+\frac{3\varepsilon(x^*)^3}{4}=0,$$

that is

$$\gamma_0=\lambda+\frac{3\varepsilon(x^*)^2}{4}. \tag{31}$$

The obtained values for γ_0 in test cases A, B and C are shown in Table 2.

Once the value of γ_0 is set, the problem to be solved will be reduced to determine u_1, from the resulting initial values problem:

$$\begin{cases}\gamma_0\left(\frac{d^2u_1}{d\tau^2}+u_1\right)=c_0\frac{\varepsilon(x^*)^3}{4}\cos 3\tau\\ u_1(0)=0\\ u_1'(0)=0\end{cases}. \tag{32}$$

A particular solution u_1 of (32) is going to be

$$u_1(\tau)=C_1\cos 3\tau+C_2\sin 3\tau.$$

Hence,

$$u_1(\tau)=-\frac{c_0}{\gamma_0}\frac{\varepsilon(x^*)^3}{32}\cos 3\tau.$$

Table 2 Values of γ_0 in tests A, B and C

	λ	ε	x^*	γ_0
A	$\frac{9}{4}$	1	1	3
B	0	1	1	$\frac{3}{4}$
C	4	-1	-1	$\frac{13}{4}$

So, the general solution u_1, of (32), will be

$$u_1(\tau) = A\cos\tau + B\sin\tau - \frac{c_0}{\gamma_0}\frac{\varepsilon(x^*)^3}{32}\cos 3\tau,$$

from which, taking into account the initial conditions ($u_1(0) = 0$ e $u_1'(0) = 0$), we deduce:

$$u_1(\tau) = \frac{c_0}{\gamma_0}\frac{\varepsilon(x^*)^3}{32}\cos\tau - \frac{c_0}{\gamma_0}\frac{\varepsilon(x^*)^3}{32}\cos 3\tau. \tag{33}$$

The approximate order 1, complete solution of the problem, will be:

$$y_1(\tau) = u_1(\tau) + u_0(\tau) = \frac{c_0}{\gamma_0}\frac{\varepsilon(x^*)^3}{32}\cos\tau - \frac{c_0}{\gamma_0}\frac{\varepsilon(x^*)^3}{32}\cos 3\tau + x^*\cos\tau. \tag{34}$$

Considering $c_0 = -1$, the approximate order 1, complete solution of the problem defined by Case A, will be:

$$y_1(\tau) = \frac{1}{96}\cos 3\tau + \frac{95}{96}\cos\tau. \tag{35}$$

The calculation of the following $n \geq 1$ terms, the pairs (γ_n, u_{n+1}) will be determined a similarly way, thus obtaining (in case of convergence) sequences

$$\gamma_n \to \gamma \quad \text{and} \quad y_{n+1} = \sum_{k=0}^{n+1} u_k(\tau) \to y.$$

Assuming the convergence of the method and remembering (23), the approximate M-order IFOHAM solution of the original problem (15) and (17), will simply be:

$$x_M(t) = y_M(\sqrt{\gamma_M}t). \tag{36}$$

The methodology followed is analogous to the methodology adopted in the application of the HAM technique, see for example page 38 of [7].

3.3 Tests and Numerical Simulations

The tests and numerical simulations will be performed in MATLAB. In the implementation of the IFOHAM algorithm we will use the symbolic toolbox of the refered platform. In the numerical resolution of the test problems we will use the Runge-Kutta method associated with the MATLAB `ode45` routine. As mentioned, we will apply the IFOHAM technique to solve the initial values problem (15) and (17) by choosing for λ, ε e x^* the values indicated in Table 1, respectively, associated with cases A, B and C.

It is noteworthy that in the application of the IFOHAM we used, in all simulations, the value $c_0 = -1$ relegating for future work the analysis of the influence of this parameter on the convergence of this technique.

In Figs. 1, 2 and 3 we plot the approximate order 3 IFOHAM solutions and the corresponding RK45 numerical solutions obtained with the Runge-Kutta method. to each case tested: Case A, Case B, and Case C.

In measuring the convergence trend we will calculate, in each iterate, the discrete version of the quadratic mean of the residue over a complete period of the approximate order M solution (discrete squared residual):

$$E_M = \frac{\int_0^{2\pi} (N\left[\gamma_M, y_M\right])^2 d\tau}{2\pi}, \tag{37}$$

which will be computed as

$$E_M \approx \frac{\sum_{k=0}^{k=N} (N\left[\gamma_M, y_M(\tau_k)\right])^2}{N+1}, \tag{38}$$

where

$$\tau_k = \frac{2k\pi}{N} \quad \text{and} \quad N = 50. \tag{39}$$

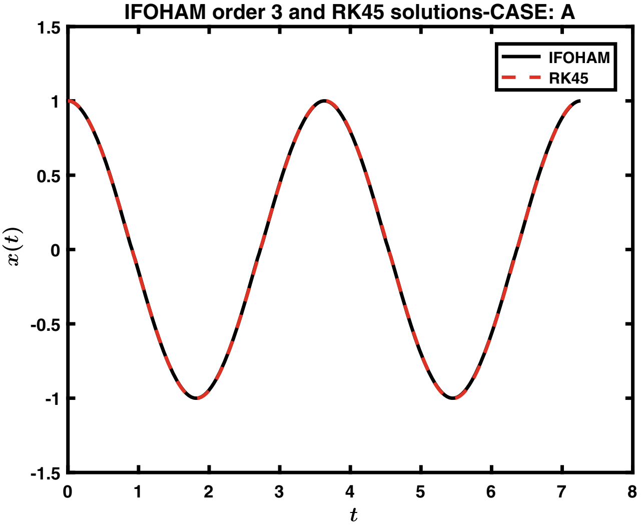


Fig. 1 Autonomous Duffing equation: approximate order 3 IFOHAM solution and RK45 numerical solution ($\gamma = 9/4$, $\varepsilon = 1$ and $x^* = 1$)

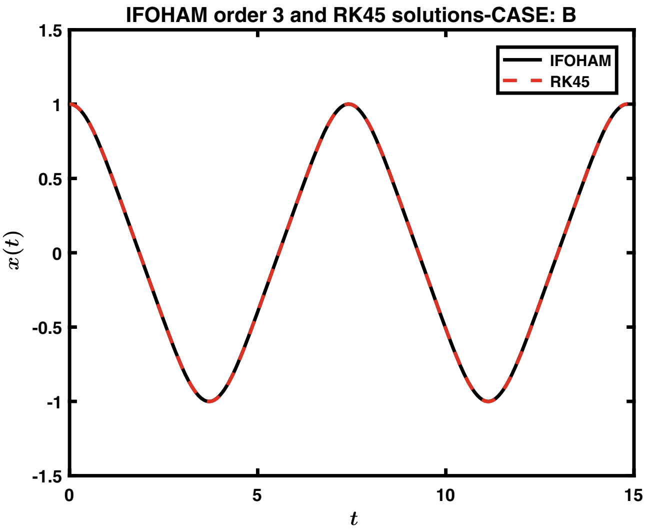


Fig. 2 Autonomous Duffing equation: approximate order 3 IFOHAM solution and RK45 numerical solution ($\gamma = 0$, $\varepsilon = 1$ and $x^* = 1$)

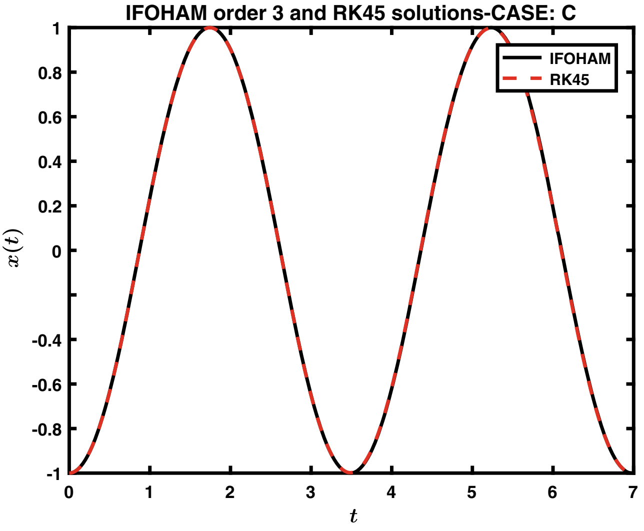


Fig. 3 Autonomous Duffing equation: approximate order 3 IFOHAM solution and RK45 numerical solution ($\gamma = 4$, $\varepsilon = -1$ and $x^* = -1$)

Table 3 Values of $\gamma_M = \omega_M^2$ (squared angular frequency) and E_M (discrete squared residual) of the first 3 iterates

IFOHAM-Case A		
Order M	γ_M	$E_M(c_0 = -1)$
0	3.000000000000000	0.031862745098039
1	2.992350260416667	0.000030759662724
2	2.992176238814215	0.000000008188297
3	2.992173084940535	0.000000000002787

Table 4 Values of $\gamma_M = \omega_M^2$ (squared angular frequency) and E_M (discrete squared residual) of the first 3 iterates

IFOHAM-Case B		
Order M	γ_M	$E_M(c_0 = -1)$
0	0.750000000000000	0.031862745098039
1	0.721354166666667	0.000492900346247
2	0.718128571631923	0.000003750953237
3	0.717803950705108	0.000000038067810

Table 5 Values of $\gamma_M = \omega_M^2$ (squared angular frequency) and E_M (discrete squared residual) of the first 3 iterates

IFOHAM-Case C		
Order M	γ_M	$E_M(c_0 = -1)$
0	3.250000000000000	0.031862745098039
1	3.242649778106509	0.000026979989273
2	3.242778638653018	0.000000004554289
3	3.242777081822536	0.000000000000845

In Tables 3, 4 and 5 we present the values obtained in the M-order approximations, using the method IFOHAM, of the squared angular frequency of the oscillation $\gamma = \omega^2$, as well as the corresponding values of the discrete squared residual E_M (38) for each one of the test cases: Case A, Case B, and Case C.

We present below the approximate order 1 solution obtained, in the resolution of case A, based on the implementation of the IFOHAM in MATLAB:

$$x_1(t) = \frac{1}{96}\cos 3\tau + \frac{95}{96}\cos\tau \quad \text{with} \quad \begin{cases} \tau = \sqrt{\gamma_1}t \\ \gamma_1 = 2.992350260416667 \end{cases} . \tag{40}$$

Of course, the expressions (35) and (40) are coincident.

Table 6 Values of $\gamma_5 = \omega_5^2$ (squared angular frequency) and E_5 (discrete squared residual) from Liao using HAM

HAM, Liao [7]

Case	γ_5	$E_5(c_0)$
A ($c_0 = -\frac{1}{3}$)	2.9921730367	1.3×10^{-14}
B ($c_0 = -\frac{4}{3}$)	0.7177741910	1.5×10^{-8}
C ($c_0 = -\frac{3}{10}$)	3.2427770978	6.7×10^{-14}

3.4 Discussion of Results

Figures 1, 2 and 3 show that the approximate order 3 solutions of the problem under study using the IFOHAM are apparently coincident with the numerical solutions resulting from the numerical simulations based on the 4th/5th order Runge-Kutta method, in all cases, A, B, and C, tested.

Looking at Tables 3, 4 and 5 we observe a convergence trend of the approximate solutions of increasing order generated by IFOHAM. This fact can be observed in the sharp decrease in the values of the quadratic mean of the residue E_M, from iterated to iterated. The decrease ratio of this parameter, from iterated to iterated, is less than 10^{-2} and can assume values of the order of 10^{-4}. Naturally, this convergence trend is accompanied by the stabilization of the values $\gamma_M = \omega_M^2$. Necessarily, the approximate values of the frequencies ω_M generated by the IFOHAM, will be all the more satisfactory the smaller the quadratic mean E_M of the corresponding residues.

In Tables 2.7 and 2.8 of [7], we can access the values of the quadratic mean of the residuals E_M and the values of $\gamma_M = \omega_M^2$ obtained using HAM, addressing the problem under study (15) and (17), under the configurations corresponding to cases A, B and C.

Note that in the referred application of HAM, it was used by Liao optimized convergence control parameters c_0. In contrast, when using IFOHAM here we used the constant value convergence control parameter $c_0 = -1$.

A rough comparison between the tabulated values in [7] and values obtained here and displayed in Tables 3, 4 and 5 suggests that the approximate 3rd order solutions obtained using IFOHAM are substantially less accurate than the approximate 5th order solutions obtained using HAM [see Table 6].

It should be noted, however, that the IFOHAM and HAM are intrinsically distinct. The IFOHAM is iterative while the HAM is constructive, facts that may explain differences in the performance of these techniques. In the computational implementation of the IFOHAM algorithm, the automatic obtaining of higher order approximations presents the difficulties normally associated with iterative processes in which the extension/complexity of each iterate significantly increases.

4 Van der Pol Equation

The van der Pol equation was introduced and studied in 1920 by the Dutch physicist Balthazar van der Pol in the context of modeling the oscillatory behavior of electric current in a triode type electronic valve [12].

Consider the autonomous van der Pol equation

$$\frac{d^2x}{dt^2} + \mu(x^2 - 1)\frac{dx}{dt} + x = 0, \tag{41}$$

where μ accounts for the intensity of nonlinear damping.

Assuming $\mu > 0$, the dynamics described by the autonomous van der Pol equation (41) is that of a self-excited system which evolves to a stable periodic behavior, that is, there is a single limit cycle (stable) for which the solution evolves regardless of the initial conditions [1].

The oscillatory solution has an amplitude close to 2. The exact value of the amplitude depends on the parameter μ [9].

4.1 Strategy for Determining the Amplitude and Frequency Using IFOHAM

The strategy we will adopt to determine the amplitude and frequency of the self-excited oscillatory response of (41) extends the adopted steps in determining the frequency of the oscillating solution of the Duffing autonomous equation in the previous section and follow the strategy defined by Liao in [3] in solving the same problem using HAM.

As noted, the van der Pol equation has a single limit cycle for which the oscillatory solution evolves independently of the initial conditions. This limit cycle is characterized by an amplitude response a and natural frequency ω. Suppose then, and without loss of generality, that $x(0)$ assumes precisely the amplitude value a (for now unknown):

$$x(0) = a. \tag{42}$$

To close the problem postulate additionally

$$\frac{dx}{dt}(0) = 0. \tag{43}$$

Then, assume the variable transformation $\tau = \omega t$ and set

$$x(t) = ay(\tau) \tag{44}$$

to highlight, in the new problem, the unknown oscillatory amplitude a and the fundamental frequency ω. Because

$$\frac{dx}{dt} = a\omega \frac{dy}{d\tau} \tag{45}$$

and

$$\frac{d^2x}{dt^2} = a\omega^2 \frac{d^2y}{d\tau^2} \tag{46}$$

the expression (41) can be rewritten in the equivalent form

$$\omega^2 \frac{d^2y}{d\tau^2} + \mu(a^2y^2 - 1)\omega \frac{dy}{d\tau} + y = 0, \tag{47}$$

with the initial conditions now associated with the new function $y = y(\tau)$, which is solution of the problem (47):

$$y(0) = 1 \quad \text{and} \quad \frac{dy}{d\tau}(0) = 0. \tag{48}$$

Once determined the function $y = y(\tau)$, solution of the differential equation (47), the sought solution $x = x(t)$, of the differential equation (41), will simply be

$$x = x(t) = ay(\omega t). \tag{49}$$

4.2 Using the IFOHAM Algorithm

Define the operator N for a condensed description of the problem. (47):

$$N\left[a, \omega, y\right] = \omega^2 \frac{d^2y}{d\tau^2} + \mu(a^2y^2 - 1)\omega \frac{dy}{d\tau} + y. \tag{50}$$

In applying the IFOHAM technique we will follow procedures common to those used in applying the HAM technique regarding the establishment of the solution expression construction rule, choosing the linear operator $\mathscr{L}$ and choosing the initial guess $u_0 = u_0(\tau)$ of the solution sought that must meet the initial conditions. These procedures are justified in Chapter 2 of [7], as well as, for example, in [5] or [3]. Then

$$\begin{cases} u_0(\tau) = \cos\tau \\ \omega_n^2 \mathscr{L}\left[u_{n+1}(\tau)\right] = c_0\left[N\left[a_n, \omega_n, \sum_{k=0}^{n} u_k(\tau)\right]\right], \quad n \geq 0 \\ \text{with} \quad u_k(\tau_0) = 0 \quad \text{and} \quad u_k'(\tau_0) = 0 \quad \forall k \in \mathscr{N} \\ y_n = \sum_{k=0}^{n} u_k(\tau) \end{cases}, \tag{51}$$

with

$$\mathscr{L}[f] = \frac{d^2 f}{d\tau^2} + f. \tag{52}$$

Note that $u_0(\tau) = \cos\tau$ satisfies the postulated initial conditions (48).

The procedure adopted to determine the sequence of iterates a_n and ω_n in the process (51) followed the approach taken and justified, for example in [5]. This procedure is based on the elimination in each iterate of the so-called secular terms $\tau\cos\tau$ and $\tau\sin\tau$ resulting from the resolution of the linear differential equation:

$$\omega_n^2 \mathscr{L}\left[u_{n+1}(\tau)\right] = c_0 \left[N\left[a_n, \omega_n, \sum_{k=0}^{n} u_k(\tau)\right]\right]. \tag{53}$$

The origin of this procedure, as referred to, and cited in [5], dates back to works by Lindstedt, Bohlin, Poincaré and Gyldén.

The described process, being convergent, will lead to the determination of sequences $\{a_n\}$, $\{\omega_n\}$ and $\{y_{n+1}\}$, such that

$$a_n \to a, \tag{54}$$

$$\omega_n \to \omega, \tag{55}$$

$$y_{n+1} = \sum_{k=0}^{n+1} u_k(\tau) \to y, \tag{56}$$

whose limits a, ω and y, satisfy the initial value problem (47) and (48). So, the determination of the sought solution (49) is accompanied by the determination of the parameters a and ω.

Assuming the convergence of the method and remembering (49), the approximate M-order IFOHAM solution of the original initial value problem (41), (42) and (43), will simply be:

$$x_M(t) = a_M y_M(\omega_M t). \tag{57}$$

4.3 Tests and Numerical Simulations

In the elaboration of tests and numerical simulations, in this section, we will continue to use MATLAB and the corresponding symbolic toolbox, as well as, the MATLAB `ode45` numerical routine. As mentioned above, the IFOHAM technique will be applied to solve the problem (41), with initial values (42) and (43).

It is noteworthy that, in the application of the IFOHAM method, we use in all simulations the value $c_0 = -1$ relegating for future work the analysis of the convergence influence of this parameter.

In Figs. 1, 2 and 3, we plot the approximate order 2 IFOHAM solutions and the RK45 numerical solutions obtained with the Runge-Kutta method, corresponding to each tested case: $\mu = 0.25$, $\mu = 0.5$ and $\mu = 1.0$.

In measuring the convergence trend of the IFOHAM method, as previously done, we will calculate in each iterate the discrete version of the quadratic mean of the residue over a complete period (discrete squared residual), of the approximate solution of order M,

$$E_M = \frac{\int_0^{2\pi} \left(N\left[a_M, \omega_M, y_M\right]\right)^2 d\tau}{2\pi}, \tag{58}$$

which will be computed as follows,

$$E_M \approx \frac{\sum_{k=0}^{k=N} \left(N\left[a_{M,}\omega_M, y_M\left(\tau_k\right)\right]\right)^2}{N+1}, \tag{59}$$

where

$$\tau_k = \frac{2k\pi}{N} \quad \text{and} \quad N = 50. \tag{60}$$

In Tables 7, 8 and 9, we present the values obtained with the IFOHAM method of the M-order discrete squared residual E_M, the angular frequency of the oscillation ω_M as well as the amplitude a_M of the oscillatory response, for each case tested: $\mu = 0.25$, $\mu = 0.5$ e $\mu = 1.0$.

Table 7 Values of E_M (discrete squared residual), ω_M (squared angular frequency) and a_M (amplitude) of the first two iterates ($\mu = 0.25$)

IFOHAM ($c_0 = -1$) - $\mu = 0.25$			
Order M	E_M	ω_M	a_M
0	0.030637254901961	1.000000000000000	2.000000000000000
1	0.001797600707410	0.996244692913550	1.997947938077153
2	0.000017619444480	0.996126327784815	2.000666070870019

Table 8 Values of E_M (discrete squared residual), ω_M (squared angular frequency) and a_M (amplitude) of the first two iterates ($\mu = 0.25$)

IFOHAM ($c_0 = -1$) - $\mu = 0.50$			
Order M	E_M	ω_M	a_M
0	0.122549019607843	1.000000000000000	2.000000000000000
1	0.027598653855655	0.986712903267367	1.990661336451152
2	0.001088915644987	0.984926844180597	2.002824407806937

Table 9 Van der Pol equation: values of E_M (discrete squared residual), ω_M (squared angular frequency) and a_M (amplitude) of the first two iterates ($\mu = 1.0$)

IFOHAM ($c_0 = -1$) - $\mu = 1.0$			
Order M	E_M	ω_M	a_M
0	0.490196078431373	1.000000000000000	2.000000000000000
1	0.408178242095261	0.970406578035910	1.947754466672885
2	0.070306106500186	0.948058574258731	2.011425421086219

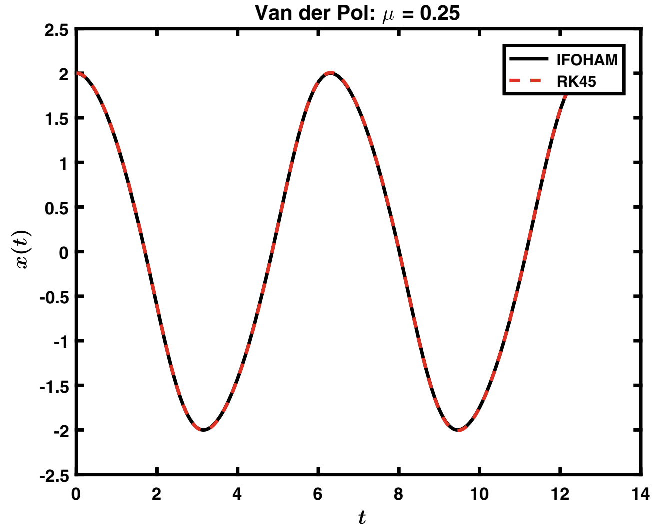


Fig. 4 Van der Pol equation: IFOHAM approximate solution of order 2 and numerical RK45 solution ($\mu = 0.25$)

4.4 Discussion of the Results

Figures 4 (case: $\mu = 0.25$) and 5 (case: $\mu = 0.5$) show that the approximate order 2 solutions of the problem under study, obtained using the IFOHAM method, appear to coincide with the numerical solutions resulting from numerical simulations based on the 4th/5th order Runge-Kutta method. In 6, where $\mu = 1.0$, clearly the approximate order 2 IFOHAM solution does not satisfactorily follow the numerical solution.

From the observation of the discrete squared residual E_M decrease in Tables 7, 8 and 9, we observe the convergence trend of the approximate increasing order solutions generated by IFOHAM.

Of course, this convergence trend is accompanied by the stabilization of the values of ω_M and a_M successively generated by the method. Necessarily, the approximate values of the frequencies ω_M and amplitudes a_M, generated by the IFOHAM method,

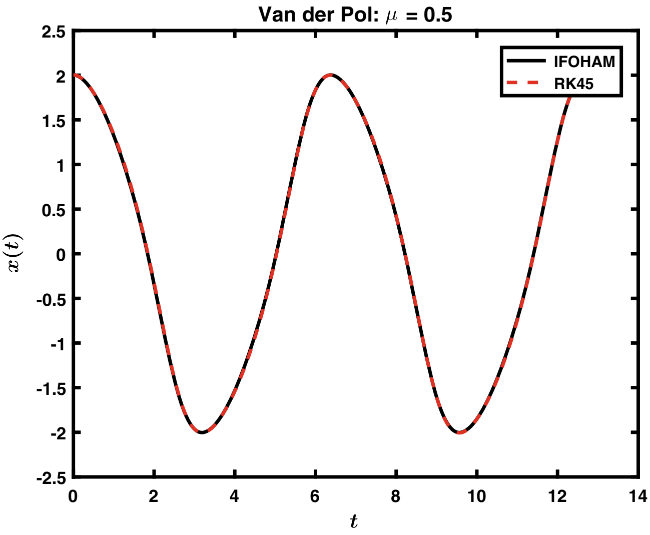


Fig. 5 Van der Pol equation: IFOHAM approximate solution of order 2 and numerical RK45 solution ($\mu = 0.5$)

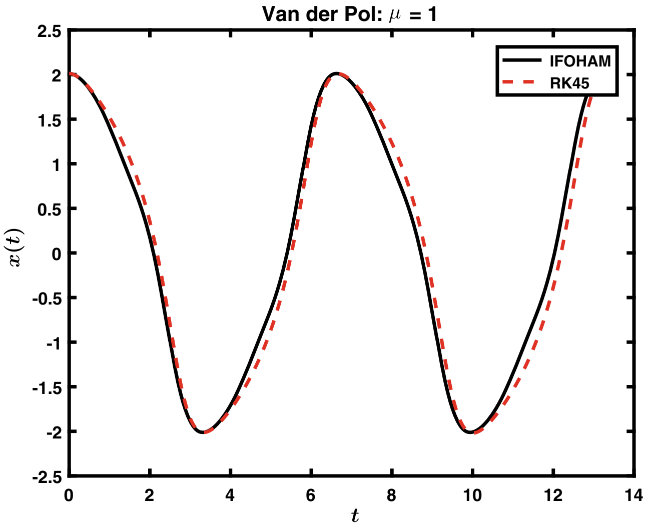


Fig. 6 IFOHAM approximate solution of order 2 and numerical RK45 solution ($\mu = 1.0$.)

will be all the more satisfactory the lower of the quadratic mean E_M of the corresponding discrete squared residual.

As μ grows, we can observe a decrease in the accuracy of the 2nd order IFOHAM approximation obtained. This explains the unsatisfactory fit between the approximate order 2 IFOHAM solution and the numerical solution shown in Fig. 6. Note that this difficulty is aggravated by the growth of the parameter μ.

Obtaining more accurate oscillatory solutions requires the calculation of approximate IFOHAM solutions of order substantially greater than 2. However, as mentioned above when discussing the results of applying this method to solve the Duffing equation, this task presents the difficulties associated with the efficiency of iterative processes in which the extent/complexity of each iterate significantly increases.

5 Conclusions and Recommendations for Future Work

5.1 Conclusions

Taking into account the results presented and the discussion made, the following conclusions are drawn:

- IFOHAM can determine periodic solutions of second order nonlinear problems and the corresponding amplitudes and oscillatory frequencies;
- IFOHAM is easy to program and apply, although, getting automatic solutions of higher order approximations presents the difficulties associated with iterative processes, in which the extension/complexity of each iterate significantly increases;
- The conjecture that IFOHAM is an extension of the Picard-Lindelöff iterative method is confirmed, which may be of theoretical interest.

5.2 Future Work

As for future work, the following developments are expected:

- Optimize the computational implementation of the IFOHAM algorithm to enable the automatic obtaining of higher order approximate solutions;
- Use IFOHAM to study the complex dynamic of the Duffing-Holmes equation;
- Use IFOHAM to study and determine the amplitude and oscillatory frequencies of coupled nonlinear systems, such as wake oscillator models;
- Study the convergence behavior of IFOHAM depending on the properties of the nonlinear differential equation under consideration, as well as, as a function of the parameter c_0.

Acknowledgements We thank the Navy Research Center (CINAV) for their financial support in carrying out this work.

References

1. Hafeez, H.Y., Ndikilar, C.E., Isyaku, S.: Analytical study of the van der pol equation in the autonomous regime. Prog. Phys. **3**, 252–255 (2015)
2. Kavacic, I., Brennan, M.J. (eds.): The Duffing Equation, Nonlinear Oscillators and Their Behaviour. Wiley, New York (2011)
3. Liao, S.-J.: An analytic approximate approach for free oscillations for self-excited systems. Int. J. Non-linear Mech. **39**, 271–280 (2004)
4. Liao, S.: The proposed homotopy analysis technique for the solution of nonlinear problems. Ph.D. thesis, Shangai Jiao Tong University, Shangai, China (1992)
5. Liao, S.: An analytic approximate approach for free oscillations of self-excited systems. Int. J. Nonlinear Mech. **39**, 271–280 (2004)
6. Liao, S.: Beyond Perturbation - Introduction to the Homotopy Analysis Method. Chapman & Hall/CRC, Boca Raton (2004)
7. Liao, S.: Homotopy Analysis Method in Nonlinear Differential Equations. Springer, Heidelberg (2012)
8. Liao, S.: Advances in Homotopy Analysis Method. World Scientific, Singapore (2014)
9. López, J.L., Abbasbandy, S., López-Ruiz, R.: Formulas for the amplitude of the van der Pol limit cycle. ArXiv, June 2008
10. Moreira, M.: IFOHAM-an iterative algorithm based on the first-order equation of HAM: exploratory preliminary results. Arxiv (2017)
11. Radhika, T.S.L., Iyengar, T. K. V., Raja Rani, T.: Approximate analytical methods for solving ordinary differential equations (2015)
12. van der Pol, B.: LXXXV. On oscillation hysteresis in a triode generator with two degrees of freedom, April 1922

Well-Posedness of Volterra Integro-Differential Equations with Fractional Exponential Kernels

N. A. Rautian

Abstract Well-defined solvability of initial boundary value problems for integro-differential equations with unbounded operator coefficients in Hilbert spaces is established in weighted Sobolev spaces on the positive semi-axis. The principal part of such equations is an abstract hyperbolic equation perturbed by terms with Volterra integral operators. These equations can be regarded as an abstract generalization of the Gurtin-Pipkin integro-differential equation that describes heat transfer in materials with memory and has a number of other applications. Numerous problems of hereditary mechanics and thermal physics have motivated the study of such equations.

1 Introduction

We study integro-differential equations with unbounded operator coefficients in Hilbert space. Let us list some problems the study of which leads to equations with the same abstract operator form as that of the equations considered in this paper. For example we consider the Gurtin-Pipkin integro-differential equation

$$u_{tt}(x,t) = u_{xx}(x,t) - \int_0^t K(t-\tau)u_{xx}(x,\tau)d\tau + f(x,t)$$

which describes the process of heat propagation in media with memory (See [3–5]), process of wave propagation in the viscoelastic media (see [1, 2]) and also arising in the problems of porous media (See [6, 7]). Moreover the Gurtin-Pipkin equation arises in the averaging procedure of a two-phased medium containing two liquids in the theory of strongly nonhomogeneous media. The kernel function $K(t)$ is the strictly positive nonincreasing function characterizing memory of media.

In this connection, it is preferable to consider integro-differential equations with operator coefficients in a Hilbert space (abstract integro-differential equations), which

N. A. Rautian (✉)
Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991,
Russian Federation
e-mail: nrautian@mail.ru

Lomonosov Moscow State University, Moscow Center for Fundamental and Applied
Mathematics, Moscow, Russian Federation

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_39

can be realized as integro-differential equations with partial derivatives with respect to the space variables where necessary. In the proof of the existence theorems, we efficiently utilize the Hilbert structure of the spaces $W^2_{2,\gamma}(\mathbb{R}_+, A)$, $L_{2,\gamma}(\mathbb{R}_+, H)$ and the Paley–Wiener theorem.

In our previous papers [8–13], the case in which the kernel K(t) can be represented by a series in decaying exponentials with positive coefficients problem (1)–(3) was studied in detail. Our approach to the study was based on the spectral analysis of the operator function which also permits proving the well-posed solvability and representing the solution of this problem in the form of a series in exponentials corresponding to points of the spectrum of the operator function $L(\lambda)$. Note that the results of [8, 9, 12, 13] were summarized in Chap. 3 of the monograph [10].

2 Statement of the Main Results

Let H be a separable Hilbert space, and let A be a self-adjoint positive operator $A^* = A$ on H with compact inverse.

$$\frac{d^2u}{dt^2} + A^2u - \int_0^t K\,(t-s)A^2u\,(s)\,ds = f\,(t)\,, \quad t \in \mathbb{R}_+, \tag{1}$$

$$u(+0) = \varphi_0, \tag{2}$$

$$u^{(1)}(+0) = \varphi_1. \tag{3}$$

The scalar function $K\,(t)$ is representable as

$$K\,(t) = \sum_{j=1}^{\infty} c_j R_j\,(t)\,, \tag{4}$$

where $c_j > 0$, $j \in \mathbb{N}$, $R_j\,(t)$ are fractional exponential Rabotnov functions (see [2], Ch. I) of the form

$$R_j(t) = t^{\alpha-1}\sum_{n=0}^{\infty}\frac{(-\beta_j)^n t^{n\alpha}}{\Gamma[(n+1)\alpha]}, \quad 0 < \alpha \le 1, \tag{5}$$

$\Gamma(\cdot)$ is the Euler gamma function. We assume that the sequence $\{\beta_j\}$ satisfies the following conditions: $0 < \beta_j < \beta_{j+1}$, $j \in \mathbb{N}$, $\beta_j \to +\infty$, $j \to +\infty$. In addition, we assume that

$$\sum_{j=1}^{\infty}\frac{c_j}{\beta_j} < 1. \tag{6}$$

The Laplace transform of $R_j(t)$ has the form

$$\hat{R}_j(\lambda) = \frac{1}{\lambda^\alpha + \beta_j},$$

(see [2], Ch. I). In this case, λ^α $(0 < \alpha \leq 1)$ is understood as the main branch of the multivalued function $f(\lambda) = \lambda^\alpha$, $\lambda \in \mathbb{C}$ with a cut along the negative real half-line: $\lambda^\alpha = |\lambda|^\alpha e^{i\alpha \arg \lambda}$, $-\pi < \arg \lambda < \pi$. Applying the inverse Laplace transform to the main branch of the multivalued function $\hat{R}_j(\lambda)$ we obtain (see [2], Ch. I) the following integral representation of function $R_j(t)$:

$$R_j(t) = \frac{1}{2\pi i} \lim_{R\to+\infty} \int\limits_{\gamma-iR}^{\gamma+iR} \frac{e^{\lambda t} d\lambda}{\lambda^\alpha + \beta_j} = \frac{\sin \pi\alpha}{\pi} \int\limits_0^{+\infty} \frac{e^{-t\tau} d\tau}{\tau^\alpha + 2\beta_j \cos \pi\alpha + \beta_j^2 \tau^{-\alpha}}.$$

Let $W_{2,\gamma}^n(\mathbb{R}_+, A^n)$ denote the Sobolev space of vector functions on the half-line $\mathbb{R}_+ = (0, \infty)$ with values in H equipped with the norm

$$\|u\|_{W_{2,\gamma}^n(\mathbb{R}_+, A^n)} \equiv \left(\int_0^\infty e^{-2\gamma t} \left(\left\|u^{(n)}(t)\right\|_H^2 + \left\|A^n u(t)\right\|_H^2 \right) dt \right)^{1/2}, \quad \gamma \geq 0.$$

See [14], Ch.1 for more details on the spaces $W_{2,\gamma}^n\left(\mathbb{R}_+, A^2\right)$. For $n = 0$, we put $W_{2,\gamma}^0\left(\mathbb{R}_+, A^0\right) = L_{2,\gamma}(\mathbb{R}_+, H)$, where $L_{2,\gamma}(\mathbb{R}_+, H)$ denotes the space of measurable functions with values in H equipped with the norm

$$\|f\|_{L_{2,\gamma}(\mathbb{R}_+, H)} = \left(\int\limits_0^{+\infty} e^{-2\gamma t} \|f(t)\|^2{}_H dt \right)^{1/2}.$$

Definition 1. A vector-function u is said to be a strong solution of problem (1)–(3) if it belongs to $W_{2,\gamma}^2(\mathbb{R}_+, A^2)$ for some $\gamma \geqslant 0$, satisfies (1) almost everywhere on the half-line $\mathbb{R}_+$, and satisfies initial condition (2), (3).

Definition 2. A vector-function u is said to be a generalized solution of problem (1)–(3) if it belongs $W_{2,\gamma}^1(\mathbb{R}_+, A)$, for some $\gamma \geqslant 0$, satisfies initial condition (2) and the identity

$$\left\langle A \left[u(t) - \int\limits_0^t K(t-s)u(s)ds \right], Av(t) \right\rangle_{L_{2,\gamma}(\mathbb{R}_+, H)} - \left\langle u'(t), v'(t) \right\rangle_{L_{2,\gamma}(\mathbb{R}_+, H)}$$
$$+2\gamma \left\langle u'(t), v(t) \right\rangle_{L_{2,\gamma}(\mathbb{R}_+, H)} = \langle f(t), v(t) \rangle_{L_{2,\gamma}(\mathbb{R}_+, H)} + (\varphi_1, v(0))_H \quad (7)$$

for some $\gamma \geqslant 0$ for all $v(t) \in W^1_{2,\gamma}(\mathbb{R}_+, A)$ satisfying the condition $\lim\limits_{t\to+\infty} v(t)e^{-2\gamma t} = 0$.

We turn the domain $Dom(A^\beta)$ of the operator $A^\beta, \beta > 0$, into a Hilbert space H_β by introducing the norm $\|\cdot\|_\beta = \|A^\beta \cdot\|$ on $Dom(A^\beta)$, which is equivalent to the norm of the graph A^β.

The following theorem gives sufficient conditions for the correct solvability of problem (1)–(3).

Theorem 1. *Assume that $Af(t) \in L_{2,\gamma_0}(\mathbb{R}_+, H)$ for some $\gamma_0 > 0$, the kernel $K(t)$ is representable as (4), (5) with a constant α $(0 < \alpha < 1)$, and condition (6) hold; in addition $\varphi_0 \in H_3$ and $\varphi_1 \in H_2$. Then there exists such $\gamma_1 > \gamma_0$, that, for all $\gamma \geqslant \gamma_1$ problem (1)–(3) has a unique solution in the space $W^2_{2,\gamma}\left(\mathbb{R}_+, A^2\right)$, satisfying the inequality*

$$\|u\|_{W^2_{2,\gamma}(\mathbb{R}_+,A^2)} \leqslant d\left(\|Af\|_{L_{2,\gamma}(\mathbb{R}_+,H)} + \left\|A^3\varphi_0\right\|_H + \left\|A^2\varphi_1\right\|_H\right), \tag{8}$$

with a constant d independent of the vector function f and the vectors φ_0 and φ_1.

It should be noticed that our approach to the proof of the theorem about the well-defined solvability of the initial boundary value problem for the abstract Gurtin–Pipkin equation essentially differs from that used by Pandolfi in [14]. Moreover, in contrast to the results of the present paper, Pandolfi studies the solvability in functional spaces defined on a finite interval $(0, T)$ of the time variable t whereas we consider the existence of solutions in weighted Sobolev spaces $W^2_{2,\gamma}(\mathbb{R}_+, A)$ on the semi-axis $\mathbb{R}_+$. In the proof of the existence theorem, we efficiently use the Hilbert structure of the spaces $W^2_{2,\gamma}(\mathbb{R}_+, A)$, $L_{2,\gamma}(\mathbb{R}_+, H)$ and the Paley–Wiener theorem.

Theorem 2. *Assume that $f(t) \in L_{2,\gamma_0}(\mathbb{R}_+, H)$ for some $\gamma_0 > 0$, the kernel $K(t)$ is representable as (4), (5) with a constant α $(0 < \alpha < 1)$, and condition (6) holds; in addition, $\varphi_0 \in H_2$ and $\varphi_1 \in H$. Then there exists $\gamma_1 > \gamma_0$ such that, for all $\gamma \geqslant \gamma_1$ problem (1)–(3) has a unique generalized solution in the space $W^1_{2,\gamma}(\mathbb{R}_+, A)$ satisfying the inequality*

$$\|u\|_{W^2_{2,\gamma}(\mathbb{R}_+,A^2)} \leqslant d\left(\|f\|_{L_{2,\gamma}(\mathbb{R}_+,H)} + \left\|A^2\varphi_0\right\|_H + \|A\varphi_1\|_H\right), \tag{9}$$

with a constant d independent of the vector function f and the vectors φ_0 and φ_1.

Considering the Laplace transform of (1) with homogeneous initial conditions, we arrive at the equation $L(\lambda)\hat{u}(\lambda) = \hat{f}(\lambda)$, where the operator function

$$L(\lambda) = \lambda^2 I + A^2 - \hat{K}(\lambda)A^2, \tag{10}$$

is the symbol of this equation, while $\hat{u}(\lambda)$ and $\hat{f}(\lambda)$ are the Laplace transforms of the vector functions $u(t)$ and $f(t)$, respectively; here, $\hat{K}(\lambda)$ is the Laplace transform of the kernel $K(t)$, which is representable as

$$\hat{K}(\lambda) = \sum_{j=1}^{\infty} \frac{c_j}{\lambda^\alpha + \beta_j}, \quad 0 < \alpha \le 1. \tag{11}$$

3 Proofs of the Theorems 1 and 2

3.1 Proof of the Auxiliary Statements

To prove Theorems 1, 2 we need the following assertions.

Proposition 1. *For any $\gamma \geqslant \gamma_1 > 0$ and all $n \in \mathbb{N}$ there exist positive constants d_1 and d_2 such that the inequalities*

$$\sup_{\operatorname{Re}\lambda > \gamma} \left| \frac{a_n}{l_n(\lambda)} \right| \leqslant d_1 < \infty, \quad \sup_{\operatorname{Re}\lambda > \gamma} \left| \frac{\lambda}{l_n(\lambda)} \right| \leqslant d_2 < \infty. \tag{12}$$

hold in the half-plane $\{\lambda : \operatorname{Re}\lambda > \gamma\}$.

Proof (Proposition 1). Let $\lambda = x + iy = |\lambda| (\cos\varphi + i \sin\varphi)$. Note that $\operatorname{sgn} y = \operatorname{sgn} \sin(\alpha\varphi)$ for $0 < \alpha \leqslant 1$. The Laplace transform $\hat{K}(\lambda)$ of the kernel $K(t)$ admits the representation

$$\hat{K}(\lambda) = \sum_{j=1}^{\infty} \left[c_j \frac{\left(|\lambda|^\alpha \cos(\alpha\varphi) + \beta_j\right) - i|\lambda|^\alpha \sin(\alpha\varphi)}{\left(|\lambda|^\alpha \cos(\alpha\varphi) + \beta_j\right)^2 + \left(|\lambda|^\alpha \sin(\alpha\varphi)\right)^2} \right]. \tag{13}$$

Consider the scalar functions

$$\mathrm{M}_n(\lambda) = \frac{l_n(\lambda)}{a_n^2} = \frac{1}{a_n^2}(L(\lambda) e_n, e_n) = \frac{\lambda^2}{a_n^2} + 1 - \sum_{k=1}^{\infty} \frac{c_k}{\lambda^\alpha + \beta_k}, \quad n \in \mathbb{N}$$

and separate their real and imaginary parts

$$\operatorname{Re}\mathrm{M}_n(\lambda) = \frac{x^2 - y^2}{a_n^2} + 1 - \operatorname{Re}\hat{K}(\lambda), \quad \operatorname{Im}\mathrm{M}_n(\lambda) = \frac{2xy}{a_n^2} - \operatorname{Im}\hat{K}(\lambda).$$

We divide the right half-plane $\{\lambda : \operatorname{Re}\lambda > \gamma_0\}$ into two the domains

$$\Omega_1 = \{|y| > \operatorname{Re}\lambda := x > \gamma_0,\ y = \operatorname{Im}\lambda\}, \quad \Omega_2 = \{\lambda : \operatorname{Re}\lambda = x > |y|, y = \operatorname{Im}\lambda\}.$$

First, we estimate the expression $\left| \dfrac{a_n}{l_n(\lambda)} \right|$ in the domain Ω_1. We have

$$\frac{|l_n(\lambda)|}{a_n^2} \geqslant |\operatorname{Im} M_n(\lambda)| = \left| \frac{2xy}{a_n^2} + \sum_{k=1}^{\infty} c_k \frac{|\lambda|^{\alpha} \sin(\alpha\varphi)}{|\lambda|^{2\alpha} + 2|\lambda|^{\alpha}\beta_k \cos(\alpha\varphi) + \beta_k^2} \right|$$

$$\geqslant \frac{2x|y|}{a_n^2} + \sum_{k=1}^{\infty} c_k \frac{|y|^{\alpha}| \left|\sin\left(\frac{\pi}{4}\alpha\right)\right|}{(|\lambda|^{\alpha} + \beta_k)^2} \geqslant \frac{2\gamma|y|}{a_n^2} + c_1 \frac{|y|^{\alpha} \left|\sin\left(\frac{\pi}{4}\alpha\right)\right|}{\left(\left(\sqrt{2}|y|\right)^{\alpha} + \beta_1\right)^2}$$

$$\geqslant \frac{2\gamma|y|\left(\left(\sqrt{2}|y|\right)^{\alpha} + \beta_1\right)^2 + c_1|y|^{\alpha} a_n^2 \sin^2\left(\frac{\pi}{4}\alpha\right)}{a_n^2\left(\left(\sqrt{2}|y|\right)^{\alpha} + \beta_1\right)^2}$$

$$\geqslant \frac{\sqrt{2\gamma|y|}\left(\left(\sqrt{2}|y|\right)^{\alpha} + \beta_1\right)\sqrt{c_1}|y|^{\alpha/2} a_n \left|\sin\left(\frac{\pi}{4}\alpha\right)\right|}{a_n^2\left(\left(\sqrt{2}|y|\right)^{\alpha} + \beta_1\right)^2}$$

$$\frac{\sqrt{2\gamma c_1}|y|^{\frac{\alpha+1}{2}} \left|\sin\left(\frac{\pi}{4}\alpha\right)\right|}{a_n\left(\left(\sqrt{2}|y|\right)^{\alpha} + \beta_1\right)} \geqslant \frac{\sqrt{2\gamma c_1}\left|\sin\left(\frac{\pi}{4}\alpha\right)\right|}{\left(\left(\sqrt{2}\right)^{\alpha} + \frac{\beta_1}{\gamma^{\alpha}}\right)} \frac{|y|^{\frac{1-\alpha}{2}}}{a_n}$$

$$\geqslant \frac{\sqrt{2\gamma c_1}\left|\sin\left(\frac{\pi}{4}\alpha\right)\right|}{\left(\left(\sqrt{2}\right)^{\alpha} + \frac{\beta_1}{\gamma^{\alpha}}\right)} \frac{\gamma^{\frac{1-\alpha}{2}}}{a_n} = \frac{k(\alpha,\gamma)}{a_n},$$

where $k(\alpha, \gamma)$ is a positive constant depending on the parameters α $(0 < \alpha < 1)$ and $\gamma > 0$. Therefore, the estimate $\frac{a_n}{|l_n(\lambda)|} \leqslant \frac{1}{k(\alpha,\gamma)}$ holds for all $\lambda \in \Omega_1$.

(2) Now let us estimate the expression $\left|\frac{a_n}{l_n(\lambda)}\right|$ in the domain Ω_2. We use condition (6) to obtain

$$\frac{|l_n(\lambda)|}{a_n^2} \geqslant |\operatorname{Re} M_n(\lambda)|$$

$$= \left| \frac{x^2 - y^2}{a_n^2} + 1 - \sum_{k=1}^{\infty} c_k \frac{|\lambda|^{\alpha}\cos(\alpha\varphi) + \beta_k}{(|\lambda|^{\alpha}\cos(\alpha\varphi) + \beta_k)^2 + (|\lambda|^{\alpha}\sin(\alpha\varphi))^2} \right| \geqslant 1 - \sum_{k=1}^{\infty} \frac{c_k}{\beta_k} > 0.$$

Indeed, note that

$$\sum_{k=1}^{\infty} c_k \frac{|\lambda|^{\alpha}\cos(\alpha\varphi) + \beta_k}{(|\lambda|^{\alpha}\cos(\alpha\varphi) + \beta_k)^2 + (|\lambda|^{\alpha}\sin(\alpha\varphi))^2}$$

$$\leqslant \sum_{k=1}^{\infty} \frac{c_k}{|\lambda|^{\alpha}\cos(\alpha\varphi) + \beta_k} \leqslant \sum_{k=1}^{\infty} \frac{c_k}{\beta_k} < 1. \tag{14}$$

holds for all $\varphi \in (-\pi/4, \pi/4)$. Therefore, for all $\lambda \in \Omega_2$ we have the estimate $\dfrac{a_n}{|l_n(\lambda)|} \leqslant \dfrac{1}{a_1\left(1 - \sum\limits_{k=1}^{\infty} \dfrac{c_k}{\beta_k}\right)}$. Thus, we finally obtain the estimate

$$\sup_{\operatorname{Re}\lambda>\gamma} \frac{a_n}{|l_n(\lambda)|} \leqslant \left(\min\left\{k(\alpha,\gamma), a_1\left(1-\sum_{k=1}^{\infty}\frac{c_k}{\beta_k}\right)\right\}\right)^{-1} =: d_1.$$

Let us estimate the expression $\left|\dfrac{\lambda}{l_n(\lambda)}\right|$ for all $\lambda = x + iy, x > \gamma$. We have

$$\begin{aligned}\left|\frac{l_n(\lambda)}{\lambda}\right| &= \left|\lambda + \frac{a_n^2}{\lambda}\left(1-\sum_{k=1}^{\infty}\frac{c_k}{\lambda^\alpha+\beta_k}\right)\right|\\ = &\left|x+iy+\frac{a_n^2(x-iy)}{x^2+y^2}\left(1-\sum_{k=1}^{\infty}c_k\frac{|\lambda|^\alpha\cos(\alpha\varphi)+\beta_k-i|\lambda|^\alpha\sin(\alpha\varphi)}{(|\lambda|^\alpha\cos(\alpha\varphi)+\beta_k)^2+(|\lambda|^\alpha\sin(\alpha\varphi))^2}\right)\right|\\ &\geqslant x+\frac{a_n^2x}{x^2+y^2}\left(1-\sum_{k=1}^{\infty}c_k\frac{|\lambda|^\alpha\cos(\alpha\varphi)+\beta_k}{(|\lambda|^\alpha\cos(\alpha\varphi)+\beta_k)^2+(|\lambda|^\alpha\sin(\alpha\varphi))^2}\right)\\ &\geqslant x+\frac{a_n^2x}{x^2+y^2}\left(1-\sum_{k=1}^{\infty}\frac{c_k}{\beta_k}\right) > x > \gamma.\end{aligned}$$

This implies the estimate $\left|\dfrac{\lambda}{l_n(\lambda)}\right| < \dfrac{1}{\gamma} =: d_2$. for all $\lambda = x + iy, x > \gamma$. The proof of the Proposition 1 is complete.

In turn, inequalities (12), (14), respectively, imply the estimates

$$\sup_{\operatorname{Re}\lambda>\gamma} \left\|AL^{-1}(\lambda)\right\| \leqslant d_1 < \infty. \tag{15}$$

$$\sup_{\operatorname{Re}\lambda>\gamma} \left\|\lambda L^{-1}(\lambda)\right\| \leqslant d_2 < \infty. \tag{16}$$

Set

$$h(t) = \int_0^t K(t-s)A^2\left(\cos(As)\varphi_0 + A^{-1}\sin(As)\varphi_1\right)ds.$$

Proposition 2. *Let the assumptions of Theorem 1 be satisfied. Then for any $\gamma \geqslant \gamma_1 > \gamma_0$ the function h satisfies the estimate*

$$\|h(t)\|_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant d_3\left(\left\|A^2\varphi_0\right\| + \|A\varphi_1\|\right) \tag{17}$$

with a constant d_3 independent of the vectors φ_0 and φ_1.

Proof (Proposition 2). To estimate the norm of the vector function $h\,(t)$ in the space $L_{2,\gamma}\,(\mathbb{R}_+,H)$, it suffices, by the Paley–Wiener theorem, to estimate the norm of the vector function $\hat{h}\,(\lambda)$ in the Hardy space $H_2(\operatorname{Re}\lambda>\gamma,H)$. The vector function $\hat{h}\,(\lambda)$ admits the representation

$$\hat{h}\,(\lambda)=\hat{K}\,(\lambda)\left[\lambda\big(\lambda^2I+A^2\big)^{-1}A^2\varphi_0+A\big(\lambda^2I+A^2\big)^{-1}A\varphi_1\right]=$$

Therefore, its norm satisfies the relation

$$\left\|\hat{h}(\lambda)\right\|^2_{H_2(\operatorname{Re}\lambda>\gamma,H)} \tag{18}$$
$$=\sup_{x>\gamma}\int\limits_{-\infty}^{+\infty}\left\|\hat{K}(x+iy)\left[(x+iy)\big((x+iy)^2I+A^2\big)^{-1}A^2\varphi_0\right.\right.$$
$$\left.\left.+\;A\big((x+iy)^2I+A^2\big)^{-1}A\varphi_1\right]\right\|^2_H dy.$$

Let us estimate the resulting integral:

$$\left\|\hat{K}(x+iy)\left[(x+iy)\big((x+iy)^2I+A^2\big)^{-1}A^2\varphi_0+A\big((x+iy)^2I+A^2\big)^{-1}A\varphi_1\right]\right\|^2_H$$

$$\leqslant C\left|\sum_{k=1}^{\infty}\frac{c_k}{(x+iy)^{\alpha}+\beta_k}\right|^2\left(\left\|(x+iy)\big((x+iy)^2I+A^2\big)^{-1}A^2\varphi_0\right\|^2_H\right.$$
$$\left.+\left\|A\big((x+iy)^2I+A^2\big)^{-1}A\varphi_1\right\|^2_H\right)$$

$$\leqslant C\left(\sum_{k=1}^{\infty}\frac{c_k}{|(x+iy)^{\alpha}+\beta_k|}\right)^2\left(\sum_{n=1}^{\infty}\frac{|x+iy|^2a_n^4|\varphi_{0n}|^2}{\left|(x+iy)^2+a_n^2\right|^2}+\sum_{n=1}^{\infty}\frac{a_n^4|\varphi_{1n}|^2}{\left|(x+iy)^2+a_n^2\right|^2}\right)$$

$$=C\left(\sum_{k=1}^{\infty}\frac{c_k}{|(x+iy)^{\alpha}+\beta_k|}\right)^2\left(\sum_{n=1}^{\infty}\frac{(x^2+y^2)a_n^4|\varphi_{0n}|^2}{(x^2-y^2+a_n^2)^2+4x^2y^2}\right.$$
$$\left.+\sum_{n=1}^{\infty}\frac{a_n^4|\varphi_{1n}|^2}{(x^2-y^2+a_n^2)^2+4x^2y^2}\right). \tag{19}$$

Note that

$$(x^2 - y^2 + a_n^2)^2 + 4x^2y^2 = \left(x^2 + (y - a_n)^2\right)\left(x^2 + (y + a_n)^2\right).$$

Moreover, for $x > \gamma$ we have the inequality

$$\frac{c_k}{|(x + iy)^\alpha + \beta_k|}$$
$$= \frac{c_k}{\sqrt{\left((x^2 + y^2)^{\alpha/2}\cos(\alpha\varphi) + \beta_k\right)^2 + (x^2 + y^2)^\alpha \sin^2(\alpha\varphi)}} \leqslant \frac{c_k}{\beta_k}.$$

Then, using the estimate (19) we obtain the following estimate of the integral

$$\left\|\hat{h}(\lambda)\right\|^2_{H_2(\mathrm{Re}\,\lambda > \gamma, H)}$$
$$\leqslant C\left(\sum_{k=1}^{\infty}\frac{c_k}{\beta_k}\right)^2 \sup_{x>\gamma}\sum_{n=1}^{\infty}\left(\int_{-\infty}^{+\infty}\frac{(x^2+y^2)a_n^4|\varphi_{0n}|^2}{\left(x^2+(y-a_n)^2\right)\left(x^2+(y+a_n)^2\right)}dy\right.$$
$$\left. + \int_{-\infty}^{+\infty}\frac{a_n^4|\varphi_{1n}|^2}{\left(x^2+(y-a_n)^2\right)\left(x^2+(y+a_n)^2\right)}dy\right)$$

$$\leqslant 2C\left(\sum_{k=1}^{\infty}\frac{c_k}{\beta_k}\right)^2 \sup_{x>\gamma}\sum_{n=1}^{\infty}\left(\int_0^{+\infty}\frac{a_n^4|\varphi_{0n}|^2}{\left(x^2+(y-a_n)^2\right)}dy + \int_0^{+\infty}\frac{a_n^4|\varphi_{1n}|^2}{a_n^2\left(x^2+(y-a_n)^2\right)}dy\right)$$
$$\leqslant C\frac{2\pi}{\gamma}\left(\sum_{k=1}^{\infty}\frac{c_k}{\beta_k}\right)^2\left(\sum_{n=1}^{\infty}a_n^4|\varphi_{0n}|^2 + \sum_{n=1}^{\infty}a_n^2|\varphi_{1n}|^2\right) = C\frac{2\pi}{\gamma}\left(\left\|A^2\varphi_0\right\|_H^2 + \|A\varphi_1\|_H^2\right).$$

The proof of the Proposition 2 is complete.

3.2 *Proof of the Theorem 1*

We begin the proof of the Theorem 1 in the case of homogeneous (zero) initial conditions ($\varphi_0 = \varphi_1 = 0$). We use Laplace transformation in order to prove the correct solvability of the problem (1)–(3). Now we are going to remind the base assertions that will be used later.

Definition 3. We denote by $H_2(\Re\lambda > \gamma, H)$ the Hardy space of vector-functions $\hat{f}(\lambda)$ taking values in the space H, holomorphic (analytic) in the semiplane $\{\lambda \in \mathbb{C} : \Re\lambda > \gamma \geqslant 0\}$ endowed with the norm

$$\sup_{x>\gamma}\int_{-\infty}^{+\infty}\left\|\hat{f}(x+iy)\right\|_H^2 dy < \infty, \quad (\lambda = x + iy). \tag{20}$$

We formulate well-known Paley–Wiener theorem for Hardy space $H_2(\Re\lambda > \gamma, H)$.

Theorem (Paley-Wiener)**.**

(1) The space $H_2(\Re\lambda > \gamma, H)$ coincides with the set of vector-functions (Laplace transformations) representing in the form

$$\hat{f}(\lambda) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\lambda t} f(t) dt, \tag{21}$$

with vector-function $f(t) \in L_{2,\gamma}(\mathbb{R}_+, H)$, $\lambda \in \mathbb{C}$, $\Re\lambda > \gamma \geqslant 0$.

(2) There exists unique vector-function $f(t) \in L_{2,\gamma}(\mathbb{R}_+, H)$ for arbitrary vector-function $\hat{f}(\lambda) \in H_2(\Re\lambda > \gamma, H)$ and the following inversion formula take place

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{f}(\gamma + iy) e^{(\gamma+iy)t} dy, \quad t \in \mathbb{R}_+, \quad \gamma \geqslant 0 \tag{22}$$

(3) The following equality take place for vector-function $\hat{f}(\lambda) \in H_2(\Re\lambda > \gamma, H)$ and $f(t) \in L_{2,\gamma}(\mathbb{R}_+, H)$, connected by relation (21):

$$\begin{aligned} \|\hat{f}\|^2_{H_2(\Re\lambda>\gamma,H)} &\equiv \sup_{x>\gamma} \int_{-\infty}^{+\infty} \left\| \hat{f}(x+iy) \right\|^2_H dy \\ &= \int_0^{+\infty} e^{-2\gamma t} \|f(t)\|^2_H \, dt \equiv \|f\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} \end{aligned} \tag{23}$$

The theorem formulated above is well-known for the scalar functions. However it is easily generalized for the vector-functions taking values in the separable Hilbert space.

Let us return to the proof of Theorem 1 for the zero initial data $\varphi_0 = \varphi_1 = 0$. We apply the Laplace transform to (1) and obtain the following representation for the Laplace transform of the solution of problem (1)–(3):

$$\hat{u}(\lambda) = L^{-1}(\lambda) \hat{f}(\lambda) \tag{24}$$

Let us prove the unique solvability of problem (1)–(3) in the space $W^2_{2,\gamma}\left(\mathbb{R}_+, A^2\right)$ for any $\gamma \geqslant \gamma_1 > \gamma_0$.

First, let us show that the vector function $A^2 u(t)$ belongs to the space $L_{2,\gamma}(\mathbb{R}_+, H)$. It is easily seen that

$$A^2 \hat{u}(\lambda) = A^2 L^{-1}(\lambda) \hat{f}(\lambda) = A L^{-1}(\lambda) A \hat{f}(\lambda) \tag{25}$$

By the Paley–Wiener theorem, $A\hat{f}(\lambda) \in H_2(\mathrm{Re}\,\lambda > \gamma_0; H)$, because $Af(t) \in L_{2,\gamma_0}(\mathbb{R}_+, H)$. Moreover, we have

$$\|Af\|_{L_{2,\gamma_0}(\mathbb{R}_+,H)} = \left\| A\hat{f} \right\|_{H_2(\mathrm{Re}\,\lambda>\gamma_0,H)}. \tag{26}$$

By the first estimate in (15) and formulas (25) and (26), we have

$$\begin{aligned}\left\|A^2u\right\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} &= \left\|A^2\hat{u}\right\|^2_{H_2(\operatorname{Re}\lambda>\gamma,H)}\\ = \left\|AL^{-1}(\lambda)\,A\hat{f}(\lambda)\right\|^2_{H_2(\operatorname{Re}\lambda>\gamma,H)} &\leqslant d_1^2\,\|Af\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)}\,.\end{aligned} \tag{27}$$

Thus, the vector function $A^2u\,(t)$ belongs to the space $L_{2,\gamma}\,(\mathbb{R}_+, H)$ and the following estimate holds:

$$\left\|A^2u\right\|_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant d_1\|Af\|_{L_{2,\gamma}(\mathbb{R}_+,H)}. \tag{28}$$

Now let us show that the vector function $\lambda^2\hat{u}\,(\lambda)$ also belongs to the space $H_2\,(\operatorname{Re}\lambda > \gamma, H)$. Note that $I = \lambda^2L^{-1}\,(\lambda) + \left(1 - \hat{K}\,(\lambda)\right)A^2L^{-1}\,(\lambda)$ for $\operatorname{Re}\lambda > \gamma$. Therefore, for $\operatorname{Re}\lambda > \gamma$ we have

$$\hat{f}\,(\lambda) = \lambda^2\hat{u}\,(\lambda) + \left(1 - \hat{K}\,(\lambda)\right)A^2L^{-1}\,(\lambda)\,\hat{f}\,(\lambda)\,. \tag{29}$$

By the assumptions about the function $K\,(t)$, the function $1 - \hat{K}\,(\lambda)$ is bounded and analytic in the half-plane $\{\lambda : \operatorname{Re}\lambda > \gamma\}$. Indeed, the following inequality holds:

$$\left|1 - \hat{K}(\lambda)\right| \leqslant 1 + \sum_{k=1}^{\infty}\frac{c_k}{|\lambda^\alpha + \beta_k|} \leqslant 1 + \sum_{k=1}^{\infty}\frac{c_k}{\beta_k} < 2.$$

By (6) the regularity (analyticity) follows from the uniform convergence of the series. Let us estimate the norm of the vector function $\lambda^2\hat{u}\,(\lambda)$ in the Hardy space $H_2\,(\operatorname{Re}\lambda > \gamma, H)$.

From the representation (29), inequality (15) and previous estimate we obtain

$$\begin{aligned}\left\|\lambda^2\hat{u}\,(\lambda)\right\|_{H_2(\operatorname{Re}\lambda>\gamma,H)} &\leqslant \|\hat{f}\,(\lambda)\|_{H_2(\operatorname{Re}\lambda>\gamma,H)}\\ +|1 - \hat{K}\,(\lambda)|\left\|AL^{-1}\,(\lambda)\,Af\,(\lambda)\right\|_{H_2(\operatorname{Re}\lambda>\gamma,H)} &\leqslant const\|Af\,(\lambda)\|_{H_2(\operatorname{Re}\lambda>\gamma,H)}.\end{aligned} \tag{30}$$

Thus, the Paley–Wiener theorem implies the inequality

$$\left\|\frac{d^2u}{dt^2}\right\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant d_1\,\|Af\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)}\,. \tag{31}$$

Finally, combining the estimates (28) and (31), we see that the vector function $u\,(t)$ belongs to the space $W^2_{2,\gamma}\,(\mathbb{R}_+, H)$, and the following estimate holds:

$$\|u\|_{W^2_{2,\gamma}(\mathbb{R}_+,A^2)} \leqslant d_2\|Af\|_{L_{2,\gamma}(\mathbb{R}_+,H)}. \tag{32}$$

Now consider problem (1)–(3) with inhomogeneous initial data φ_0 and φ_1. Set

$$u(t) = \cos(At)\,\varphi_0 + A^{-1}\sin(At)\,\varphi_1 + w(t)\,. \tag{33}$$

Then the vector function $w(t)$ is a solution of the problem

$$\frac{d^2w}{dt^2} + A^2 w(t) - \int_0^t K(t-s)A^2 w(s)\,ds = f_1(t)\,, \tag{34}$$

$$w(+0) = w^{(1)}(+0) = 0, \tag{35}$$

where $f_1(t) = f(t) - h(t)$ and

$$h(t) = \int\limits_0^t K(t-s)A^2\left(\cos(As)\,\varphi_0 + A^{-1}\sin(As)\,\varphi_1\right)ds.$$

To prove the theorem, it suffices to prove the inequality

$$\|Af_1\|_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant \|Af\|_{L_{2,\gamma}(\mathbb{R}_+,H)} + \|Ah\|_{L_{2,\gamma}(\mathbb{R}_+,H)} < \infty. \tag{36}$$

The estimate (17) implies the estimate

$$\|Ah(t)\|_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant d_4\left(\left\|A^3\varphi_0\right\| + \left\|A^2\varphi_1\right\|\right) \tag{37}$$

with a constant d_4 independent of the vectors φ_0 and φ_1.

Now let us prove the uniqueness of the strong solution of problem (1)–(3). Assume that there exist two distinct strong solutions $u_1(t)$ and $u_2(t)$ of problem (1)–(3). Then the vector function $v(t) = u_1(t) - u_2(t)$ is a strong solution of problem (1)–(3) with zero right-hand side $f(t) \equiv 0$ and zero initial vectors $\varphi_0 = \varphi_1 = 0$, and its Laplace transform $\hat{v}(\lambda)$ satisfies the equation $L(\lambda)\hat{v}(\lambda) = 0$. Therefore, we have $\hat{v}(\lambda) = 0$ and, by the inversion formula for the Laplace transform, $v(t) \equiv 0$. The proof of Theorem 1 is complete.

3.3 Proof of the Theorem 2

Assume that $u(t)$ is the strong solution of problem (1)–(3) with zero initial data $\varphi_0 = \varphi_1 = 0$. Applying the Laplace transform to (1), we obtain the following representation for the Laplace transform of the strong solution $u(t)$ of problem (1)–(3):

$$\hat{u}(\lambda) = L^{-1}(\lambda)\,\hat{f}(\lambda) \tag{38}$$

Consider the projection $u_n(t)$ of the vector function $u(t)$ onto the one-dimensional subspace spanned by the vector e_n; i.e., $u_n(t) = (u(t), e_n)_H$. Then we have $\hat{u}_n(\lambda) = \big(\hat{u}(\lambda), e_n\big)_H = l_n^{-1}(\lambda)\hat{f}_n(\lambda)$.

First, let us prove the generalized solvability in the space $W^1_{2,\gamma}(\mathbb{R}_+, A)$ of problem (1), (2) with zero initial data $\varphi_0 = \varphi_1 = 0$ for any $\gamma \geqslant \gamma_1 > \gamma_0$. To this end, we prove the following assertion.

Claim 1. *If a function $u_n(t)e_n$ is a strong solution of the problem*

$$\frac{d^2u}{dt^2} + A^2u - \int_0^t K(t-s)A^2u(s)\,ds = f_n(t)\,e_n, \quad t \in \mathbb{R}_+, \tag{39}$$

$$u(+0) = 0, \quad u^{(1)}(+0) = 0, \tag{40}$$

then the function $u_n(t)e_n$ is a generalized solution of this problem.

Proof. Indeed, taking the inner product of both sides of (39) by the function $v(t) \in W^1_{2,\gamma}(\mathbb{R}_+, A)$ in the space $L_{2,\gamma}(\mathbb{R}_+, H)$, where $\gamma \geqslant \gamma_1 > \gamma_0$ we obtain

$$\int_0^{+\infty} \big(u_n''(t)e_n, v(t)\big)_H e^{-2\gamma t}dt$$
$$+ \int_0^{+\infty} \left(A^2\left[u_n(t)e_n - \int_0^t K(t-s)u_n(s)e_n ds\right], v(t)\right)_H e^{-2\gamma t}dt$$
$$= \int_0^{+\infty} (f_n(t)e_n, v(t))_H e^{-2\gamma t}dt.$$

We integrate the first term by parts and obtain

$$\int_0^{+\infty} \big(u_n''(t)e_n, v(t)\big)_H e^{-2\gamma t}dt$$
$$= -(\varphi_1, v(0))_H - \big\langle u_n'(t), v'(t)\big\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)} + 2\gamma\big\langle u_n'(t), v(t)\big\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}.$$

Transforming the second term, we obtain

$$\int_0^{+\infty}\left(A^2\left[u_n(t)e_n-\int_0^t K(t-s)u_n(s)e_n ds\right], v(t)\right)_H e^{-2\gamma t}dt$$
$$=\left\langle A\left[u_n(t)e_n-\int_0^t K(t-s)u_n(s)e_n ds\right], Av(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}.$$

Thus, the function $u_n(t)e_n$ satisfies the identity

$$\left\langle A\left[u_n(t)e_n+\int_0^t K(t-s)u_n(s)e_n ds\right], Av(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}$$
$$-\left\langle u'_n(t)e_n, v'(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}+2\gamma\left\langle u'_n(t)e_n, v(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}=\langle f_n(t)e_n, v(t)\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}$$

and hence is a generalized solution of problem (39), (40). The proof of the claim is complete.

Corollary 1. *If the vector function $S_N(t)=\sum_{n=1}^N u_n(t)e_n$ is a strong solution of the problem*

$$\frac{d^2u}{dt^2}+A^2u-\int_0^t K(t-s)A^2u(s)\,ds=F_N(t), \quad t\in\mathbb{R}_+, \tag{41}$$

$$u(+0)=0, \quad u^{(1)}(+0)=0, \tag{42}$$

where $F_N(t)=\sum_{n=1}^N f_n(t)$, then the vector function $S_N(t)$ is a generalized solution of problem (41), (42)*; i.e., the following identity holds:*

$$\left\langle A\left[S_N(t)+\int_0^t K(t-s)S_N(s)ds\right], Av(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}-\left\langle S'_N(t), v'(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+)}$$
$$+2\gamma\left\langle S'_N(t), v(t)\right\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)}=\langle F_N(t), v(t)\rangle_{L_{2,\gamma}(\mathbb{R}_+,H)} \tag{43}$$

Now we return to the proof of Theorem 2. Let us show that if its conditions are satisfied, then the vector function $u(t)=\sum_{n=1}^\infty u_n(t)e_n$ (where, for each $n\in\mathbb{N}$, the function $u_n(t)e_n$ is a strong solution of the corresponding problem (39), (40)) is a generalized solution of problem (1)–(3).

To this end, using the estimate (12) we show that the sequence $S_N(t)=\sum_{n=1}^N u_n(t)e_n$ is a Cauchy sequence in the space $W^1_{2,\gamma}(\mathbb{R}_+, A)$. We have

$$\|S_N(t) - S_M(t)\|^2_{W^1_{2,\gamma}(\mathbb{R}_+,A)} = \left\| \sum_{n=M+1}^{N} u_n(t)e_n \right\|^2_{W^1_{2,\gamma}(\mathbb{R}_+,A)}$$

$$= \left\| A \sum_{n=M+1}^{N} u_n(t)e_n \right\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} + \left\| \sum_{n=M+1}^{N} u'{}_n(t)e_n \right\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)}$$

$$= \left\| A \sum_{n=M+1}^{N} \hat{u}_n(\lambda)e_n \right\|^2_{H_2(\Re\lambda>0,H)} + \left\| \sum_{n=M+1}^{N} \lambda\hat{u}_n(\lambda)e_n \right\|^2_{H_2(\Re\lambda>0,H)}$$

$$= \sup_{x>\gamma} \int_{-\infty}^{+\infty} \left(\sum_{n=M+1}^{N} |a_n \hat{u}_n(x+iy)|^2 + \sum_{n=M+1}^{N} |(x+iy)\hat{u}_n(x+iy)|^2 \right) dy$$

$$= \sup_{x>\gamma} \int_{-\infty}^{+\infty} \left(\sum_{n=M+1}^{N} \left| \frac{a_n \hat{f}_n(x+iy)}{l_n(x+iy)} \right|^2 + \sum_{n=M+1}^{N} \left| \frac{(x+iy)\hat{f}_n(x+iy)}{l_n(x+iy)} \right|^2 \right) dy$$

$$\leqslant \int_{-\infty}^{+\infty} \sum_{n=M+1}^{N} \sup_{x>\gamma} \left(\left| \frac{a_n}{l_n(x+iy)} \right|^2 + \left| \frac{x+iy}{l_n(x+iy)} \right|^2 \right) \left| \hat{f}_n(x+iy) \right|^2 dy$$

$$\leqslant d_5 \sup_{x>\gamma} \int_{-\infty}^{+\infty} \sum_{n=M+1}^{N} \left| \hat{f}_n(x+iy) \right|^2 dy = d_5 \left\| \sum_{n=M+1}^{N} \hat{f}_n(\lambda)e_n \right\|_{H_2(\Re\lambda>0,H)}$$

$$= d_5 \left\| \sum_{n=M+1}^{N} f_n(t)e_n \right\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)}. \qquad (44)$$

By the assumptions of Theorem 2, the vector function $f(t)$ belongs to the space $L_{2,\gamma}(\mathbb{R}_+, H)$, and therefore, the sequence of vector functions $S_N(t)$ converges in the space $W^1_{2,\gamma}(\mathbb{R}_+, A)$ to the vector function $u(t)$ if the sequence $F_N(t)$ converges to the vector function $F(t)$ in the space $L_{2,\gamma}(\mathbb{R}_+, H)$. Passing to the limit as $N \to +\infty$ in identity (43), we obtain identity (7) for $\varphi_1 = 0$; i.e., the function $u(t) = \sum_{n=1}^{\infty} u_n(t)e_n$ is a generalized solution of problem (1)–(3).

Now let us estimate the norm of the generalized solution $u(t) = \sum_{n=1}^{\infty} u_n(t)e_n$ in the space $W^1_{2,\gamma}(\mathbb{R}_+, A)$. By setting $M = 0$ and $S_M(t) = 0$ in the chain of inequalities (44), we obtain the inequality

$$\|S_N(t)\|^2_{W^1_{2,\gamma}(\mathbb{R}_+,A)} \leqslant d_5 \left\| \sum_{n=1}^{N} f_n(t)e_n \right\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant d_5 \, \|f(t)\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} \, .$$

from which, passing to the limit as $N \to +\infty$ we obtain the estimate

$$\|u(t)\|^2_{W^1_{2,\gamma}(\mathbb{R}_+,A)} \leqslant d_5 \, \|f(t)\|^2_{L_{2,\gamma}(\mathbb{R}_+,H)} \, . \tag{45}$$

Now consider problem (1)–(3) with inhomogeneous initial data φ_0 and φ_1. Set

$$u\,(t) = \cos\,(At)\,\varphi_0 + A^{-1}\sin\,(At)\,\varphi_1 + w\,(t)\,. \tag{46}$$

Claim 2. *The vector function $u(t)$ is a generalized solution of problem* (1)–(3) *whenever the vector function $w\,(t)$ is a generalized solution of the problem*

$$\frac{d^2w}{dt^2} + A^2w\,(t) - \int_0^t K\,(t-s)A^2w\,(s)\,ds = f_1\,(t)\,, \tag{47}$$

$$w\,(+0) = w^{(1)}\,(+0) = 0, \tag{48}$$

where $f_1\,(t) = f\,(t) - h\,(t)$,

$$h\,(t) = \int\limits_0^t K\,(t-s)A^2\left(\cos\,(As)\,\varphi_0 + A^{-1}\sin\,(As)\,\varphi_1\right)ds.$$

The conditions of Theorem 2 and Proposition imply the estimate

$$\|f_1\|_{L_{2,\gamma}(\mathbb{R}_+,H)} \leqslant \|f\|_{L_{2,\gamma}(\mathbb{R}_+,H)} + \|h\|_{L_{2,\gamma}(\mathbb{R}_+,H)} < \infty. \tag{49}$$

Thus, the assumptions of Theorem 2 are satisfied for problem (47), (48). A straightforward substitution of the vector function $u(t)$ given by (46) into identity (7) readily proves the assertion of Claim 2. Moreover, the estimate (45) and Proposition 2 imply the estimate (9). The proof of Theorem 2 is complete.

Acknowledgements This work was supported by Russian Foundation for Basic Research (grant no. 20-01-00288 A).

References

1. Il'yushin, A.A., Pobedrya, B.E.: Osnovy matematicheskoi teorii termovyazkouprugosti (Foundations of Mathematical Theory of Thermoviscoelasticity). Nauka, Moscow (1970)
2. Rabotnov, Yu.N.: Elementy nasledstvennoi mekhaniki tverdykh tel (Elements of Hereditary Mechanics of Solids). Nauka, Moscow (1977)
3. Lykov, A.V.: Problema teplo- i massoobmena (Heat and Mass Exchange Problem). Nauka i Tekhnika, Minsk (1976)
4. Gurtin, M.E., Pipkin, A.C.: Theory of heat conduction with finite wave speed. Arch. Ration. Mech. Anal. **31**, 113–126 (1968)
5. Eremenko, A., Ivanov, S.: Spectra of the Gurtin-Pipkin type equations. SIAM J. Math. Anal. **43**(5), 2296–2306 (2011). https://doi.org/10.1137/100811908
6. Gavrikov, A.A., Ivanov, S.A., Knyaz'kov, D.Yu., Samarin, V.A., Shamaev, A.S., Vlasov, V.V.: Spectral properties of combined media. J. Math. Sci. **164**(6), 948–963 (2010). https://doi.org/10.1007/s10958-010-9776-5
7. Zhikov, V.V.: On an extension of the method of two-scale convergence and its applications. Sb. Math. **191**(7), 973–1014 (2000). https://doi.org/10.1070/SM2000V191N07ABEH000491
8. Rautian, N.A., Vlasov, V.V.: Well-defined solvability and spectral analysis of abstract hyperbolic. J. Math. Sci. **179**(3), 390–415 (2011). https://doi.org/10.1007/s10958-011-0600-7
9. Rautian, N.A., Vlasov, V.V.: Properties of solutions of integro-differential equations arising in heat and mass transfer theory. Trans. Mosc. Math. Soc. 185–204 (2014). https://doi.org/10.1090/S0077-1554-2014-00231-4
10. Rautian, N.A., Vlasov, V.V.: Spektral'nyi analiz funktsional'no-differentsialnykh uravnenii (Spectral Analysis of Functional-Differential Equations). MAKS Press, Moscow (2016)
11. Rautian, N.A., Vlasov, V.V.: Well-posedness and spectral analysis of integrodifferential equations arising in viscoelasticity theory. J. Math. Sci. **233**(4), 555–577 (2018). https://doi.org/10.1007/s10958-018-3943-5
12. Rautian, N.A., Vlasov, V.V.: Well-posedness and spectral analysis of integrodifferential equations arising in viscoelasticity theory. Study of operator models arising in viscoelasticity. Sovrem. Mat. Fundam. Napravl. **64**(1), 60–73 (2018)
13. Vlasov, V.V., Wu, J.: Solvability and spectral analysis of abstract hyperbolic equations with delay. Funct. Differ. Equ. **16**(4), 751–768 (2009)
14. Lions, J.L., Magenes, E.: Nonhomogeneous Boundary-Value Problems and Applications. Springer, Heidelberg (1972)

On Non-local Boundary-Value Problems for Higher-Order Non-linear Functional Differential Equations

Nataliya Dilna

Abstract Some optimal, in a sense, general conditions sufficient for a unique solvability of the non-local boundary-value problem for higher-order non-linear functional differential equations are established. The class of equations considered covers, in particular, non-linear equations with transformed argument, integro-differential equations and neutral equations. Example is presented to illustrate the optimality of results.

Keywords Boundary-value problem · Functional-differential equations · Non-local conditions · Unique solvability · Differential inequality · Optimal conditions

MSC 2010: 34K10 · 34K38

1 Problem Formulation and Definition

The paper deals with the question on the existence and uniqueness of a solution of a non-local boundary-value problem for higher-order non-linear functional differential equations of the general form

$$u_k{}^{(m)}(t) = (f_k u)(t) + q_k(t), \quad t \in [a,b], \quad k = \overline{1,n}, \tag{1}$$

$$u_k{}^{(m-i)}(a) = \varphi_{m-ik}(u), \quad k = \overline{1,n}, \quad i = \overline{1,m}, \tag{2}$$

where $m, n \in \mathbb{N}$, $i \le m$, $f_k : \mathscr{W}^m([a,b],\mathbb{R}^n) \to \mathscr{L}_1([a,b],\mathbb{R})$, $k = \overline{1,n}$, are, generally speaking, non-linear operators, $q_k \in \mathscr{L}_1([a,b],\mathbb{R})$, $\varphi_{ik} : \mathscr{W}^m([a,b],\mathbb{R}^n) \to \mathbb{R}$, $i = \overline{1,m}$, $k = \overline{1,n}$, non-linear functionals from the space $\mathscr{W}^m([a,b],\mathbb{R}^n)$ of vector-functions with absolutely continuous coordinates $u^{(m-1)}$.

N. Dilna (✉)
Mathematical Institute of the Slovak Academy of Sciences, Štefánikova 49 St.,
814 73 Bratislava, Slovakia
e-mail: nataliya.dilna@mat.savba.sk

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_40

The investigation has been motivated, mainly, by the recent publications [4, 8, 9, 13, 16]. The general problem (1), (2) are active studied in modern literature (see, [3–5, 10] and references therein).

We have established conditions sufficient for a unique solvability of the non-local boundary-value problem for systems of non-linear higher-order functional differential equations (1), (2). The idea of proof of our results is based on the application of an abstract result ensuring the unique solvability of an equation with an operator satisfying Lipschitz-type conditions with respect to a suitable cone (see, Theorem 49.4 from [12]). The main result of this paper and corollaries are in Sect. 3, auxiliary statements one can find in Sect. 4; the proof of the main general Theorem 1 is in Sect. 5; conditions sufficient for the unique solvability of linear functional differential equations are in Sect. 6; it is shown in Sect. 7 that the obtained results are, in a sense, optimal.

In the general case, operator f from Eq. (1) is given on $\mathscr{W}^m([a,b],\mathbb{R}^n)$ only and, thus, the right-hand side term of Eq. (1) may contain terms with derivatives, and, hence, the statements presented in what follows are applicable, in particular, to functional differential equations of the neutral type.

By a *solution* of problem (1), (2), as usual (see, e.g., [1]), we mean a vector function $u = (u_k)_{k=1}^n : [a,b] \to \mathbb{R}^n$ whose components are absolutely continuous, satisfy system (1) almost everywhere on the interval $[a,b]$, and possess properties (2) at the point a for $i = \overline{1,m}$.

Definition 1. A linear operator $l = (l_k)_{k=1}^n : \mathscr{W}^m([a,b],\mathbb{R}^n) \to \mathscr{L}_1([a,b],\mathbb{R}^n)$ is said to belong to the *set* $\mathscr{S}_{\overline{h_0,h_{m-1}}}$ if the boundary value problem

$$u_k{}^{(m)}(t) = (l_k u)(t) + q_k(t), \quad t \in [a,b], \quad k = \overline{1,n}, \tag{3}$$

$$u_k{}^{(m-i)}(a) = h_{m-ik}(u) + c_{m-ik}, \quad k = \overline{1,n}, \quad i = \overline{1,m}, \tag{4}$$

where $h_{ik} : \mathscr{W}^m \to \mathbb{R}$, $i = 1,2$, $k = \overline{1,n}$, are linear functionals, has a unique solution $u = (u_k)_{k=1}^n$ for any $q_k \in \mathscr{L}_1([a,b],\mathbb{R})$ and, moreover, the solution of (3), (4) possesses the property

$$\min_{t\in[a,b]} u_k(t) \geq 0, \qquad k = \overline{1,n}, \tag{5}$$

whenever the components of the function q_k, $k = \overline{1,n}$, and constants c_{ik}, $k = \overline{1,n}$, $i = \overline{1,m}$, appearing in (3) are non-negative almost everywhere on $[a,b]$.

2 Notation

Through whole work will used the next notations.

1. We fix a bounded interval $[a,b]$ and a natural numbers n and m.
2. $\mathbb{R} := (-\infty,\infty)$; $\|x\| := \max_{1\leq i\leq n} |x_i|$ for $x = (x_i)_{i=1}^n \in \mathbb{R}^n$.

3. $\mathscr{L}_1([a,b],\mathbb{R}^n)$ is the Banach space of all the Lebesgue integrable vector-functions $u:[a,b]\to\mathbb{R}^n$ with the standard norm

$$\mathscr{L}_1([a,b],\mathbb{R}^n)\ni u\longmapsto\int_a^b\|u(s)\|\,ds.$$

4. $\mathscr{W}^m([a,b],\mathbb{R}^n)$ is set of vector-functions $u=(u_i)_{i=1}^n:[a,b]\to\mathbb{R}^n$ with $u^{(m-1)}$ absolutely continuous on $[a,b]$ and the norm given by the formula

$$\mathscr{W}^m([a,b],\mathbb{R}^n)\ni u\longmapsto\|u\|_{\mathscr{W}^m}:=\int_a^b\|u^{(m)}(s)\|\,ds+\sum_{i=0}^{m-1}\|u^{(i)}(a)\|. \qquad (6)$$

5. For $i=\overline{1,n}$, by $\mathscr{W}^m_{(0)}([a,b],\mathbb{R}^n)$ we denote the set of vector-functions $u=(u_i)_{i=1}^n:[a,b]\to\mathbb{R}^n$ from $\mathscr{W}^m([a,b],\mathbb{R}^n)$ such that the components of u are non-negative a.e. on $[a,b]$.
6. For $i=\overline{1,n}$, by $\mathscr{W}^m_{(m)}([a,b],\mathbb{R}^n)$ we denote the set of vector-functions $u=(u_i)_{i=1}^n:[a,b]\to\mathbb{R}^n$ from $\mathscr{W}^m([a,b],\mathbb{R}^n)$ such that the components of $u_i^{(m)}(t)$ are non-negative a.e. on $[a,b]$ and $u_i^{(j)}(a)\geq 0$, for $i=\overline{1,n}$, $0\leq j\leq m-1$.

In what follows, the symbols $\mathscr{W}^m([a,b],\mathbb{R}^n)$, $\mathscr{W}^m_{(0)}([a,b],\mathbb{R}^n)$, $\mathscr{W}^m_{(m)}([a,b],\mathbb{R}^n)$, corresponding to the fixed a, b, and n will usually appear simply as $\mathscr{W}^m$, $\mathscr{W}^m_{(0)}$, $\mathscr{W}^m_{(m)}$.

3 General Results

3.1 Main Theorem

The main general result of this paper is the next theorem.

Theorem 1. *Assume that there exist some linear operators $p=(p_k)_{k=1}^n:\mathscr{W}^m\to\mathscr{L}_1$, $\xi=(\xi_k)_{k=1}^n:\mathscr{W}^m\to\mathscr{L}_1$, and linear functionals $h_i=(h_{ik})_{ik=1}^n:\mathscr{W}^m\to\mathbb{R}^n$, $r_i=(r_{ik})_{k=1}^n:\mathscr{W}^m\to\mathbb{R}^n$, $i=\overline{1,m-1}$, which satisfy the inclusions*

$$p\in\mathscr{S}_{\overline{h_0,h_{m-1}}},\qquad \frac{1}{2}(p+\xi)\in\mathscr{S}_{\overline{\frac{1}{2}(h_0+r_0),\frac{1}{2}(h_{m-1}+r_{m-1})}}, \qquad (7)$$

and such that for arbitrary functions $u=(u_k)_{k=1}^n:[a,b]\to\mathbb{R}^n$, $v=(v_k)_{k=1}^n:[a,b]\to\mathbb{R}^n$ from $\mathscr{W}^m$ with the properties

$$u_k(t)\geq v_k(t),\quad t\in[a,b],\quad k=\overline{1,n}, \qquad (8)$$

the inequalities

$$\xi_k(u-v)(t) \le (f_k u)(t) - (f_k v)(t) \le p_k(u-v)(t), \quad t \in [a,b], \quad k = \overline{1,n}, \tag{9}$$

and

$$r_{ik}(u-v)(t) \le \varphi_{ik}(u) - \varphi_{ik}(v) \le h_{ik}(u-v), \quad i = \overline{1,m-1}, \quad k = \overline{1,n} \tag{10}$$

hold.

Then the non-local boundary-value problem (1), (2) *has a unique solution for an arbitrary function* $q \in \mathscr{L}_1$.

We get the next corollaries from this theorem.

3.2 Corollaries

Theorem 2. *Suppose that for arbitrary absolutely continuous vector-functions* $u = (u_k)_{k=1}^n : [a,b] \to \mathbb{R}^n$ *and* $v = (v_k)_{k=1}^n : [a,b] \to \mathbb{R}^n$ *with properties* (8) *the inequalities* (10) *and*

$$\left|(f_k u)(t) - (f_k v)(t) - l_{1k}(u-v)(t)\right| \le l_{2k}(u-v)(t), \quad k = \overline{1,n}, \tag{11}$$

hold for some linear operators $l_j = (l_{jk})_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$, $j = 1, 2$, *which satisfy inclusions*

$$l_1 + l_2 \in \mathscr{S}_{\overline{h_0, h_{m-1}}}, \qquad l_1 \in \mathscr{S}_{\overline{\frac{1}{2}(h_0+r_0), \frac{1}{2}(h_{m-1}, h_{m-1})}}. \tag{12}$$

Then the boundary-value problem (1), (2) *is uniquely solvable for an arbitrary* $q \in \mathscr{L}_1$.

Proof. Obviously, that condition (11) is equivalent to the relation

$$\begin{aligned} & l_{1k}(u-v)(t) - l_{2k}(u-v)(t) \\ & \qquad \le (f_k u)(t) - (f_k v)(t) \le l_{1k}(u-v)(t) + l_{2k}(u-v)(t) \end{aligned} \tag{13}$$

for arbitrary functions u and v from $\mathscr{W}^m$ with properties (8) and $t \in [a,b]$. Let us put for any $k = 1, 2, \ldots, n$

$$(p_k u)(t) := (l_{1k} u)(t) + (l_{2k} u)(t) \quad \text{and} \quad (\xi_k u)(t) := (l_{1k} u)(t) - (l_{2k} u)(t). \tag{14}$$

Then (13) means that f satisfies condition (9). It is also clear that (12) provides of (7) with p_k and ξ_k given by (14). Application of Theorem 1 thus leads us to the assertion of Theorem 2.

The next corollaries is true.

Definition 2. We say, that an operator $p = (p_k)_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$ is *positive*, if for arbitrary vector-function $u \in \mathscr{W}_{(0)}^m$ the next inequality

$$(p_k u)(t) \geq 0, \quad k = \overline{1, n},$$

is true for a.e. $t \in [a, b]$.

Corollary 1. *Let there exist some positive linear operators $g_i = (g_{ik})_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$, $i = 1, 2$, and linear functionals $h_{ik} : \mathscr{W}^m \to \mathbb{R}$ and $r_{ik} : \mathscr{W}^m \to \mathbb{R}$ with property* (10) *for which the inequality*

$$|(f_k u)(t) - (f_k v)(t) + g_{2k}(u - v)(t)| \leq g_{1k}(u - v)(t), \quad k = \overline{1, n}, \tag{15}$$

is true for u and v from $\mathscr{W}^m$ with property (8)*. Furthermore, assume that inclusions*

$$g_1 \in \mathscr{S}_{\overline{h_0, h_{m-1}}}, \qquad -\frac{1}{2} g_2 \in \mathscr{S}_{\overline{\frac{1}{2}(h_0 + r_0), \frac{1}{2}(h_{m-1}, h_{m-1})}} \tag{16}$$

are fulfilled.

Then the non-local boundary-value problem (1)*,* (2) *has a unique solution for an arbitrary function $q \in \mathscr{L}_1$.*

Proof. It follows from assumption (15) and the positivity of the operator g_2 that the relations

$$\begin{aligned} &|(f_k u)(t) - (f_k v)(t) + \tfrac{1}{2} g_{2k}(u - v)(t)| \\ &\qquad = |(f_k u)(t) - (f_k v)(t) + g_{2k}(u - v)(t) - \tfrac{1}{2} g_{2k}(u - v)(t)| \\ &\qquad \leq g_{1k}(u - v)(t) + \tfrac{1}{2}|g_{2k}(u - v)(t)| = g_{1k}(u - v)(t) + \tfrac{1}{2} g_{2k}(u - v)(t) \end{aligned}$$

are true for any u and v with properties (8). This means that $f = (f_k)_{k=1}^n$ admits estimate (11) with the operators l_1 and l_2 defined by the equalities

$$l_1 := -\frac{1}{2} g_2, \quad l_2 := g_1 + \frac{1}{2} g_2. \tag{17}$$

Moreover, assumption (16) guarantees that inclusions (12) hold for l_1 and l_2 from (17). Thus, we can apply Theorem 2, which leads us to the required assertion.

4 Auxiliary Propositions

To prove our main result, we use the following statement on the unique solvability of an equation with a Lipschitz type non-linearity established in [12].

Let us consider the abstract operator-equation

$$Fx = z, \tag{18}$$

where $F : E_1 \to E_2$ is a mapping between a normed space $\langle E_1, \|\cdot\|_{E_1}\rangle$ and a Banach space $\langle E_2, \|\cdot\|_{E_2}\rangle$ over the field $\mathbb{R}$, and z is an arbitrary element from E_2.

Let $K_i \subset E_i, i = 1, 2$, be cones [11]. The cones $K_i, i = 1, 2$, induce natural partial orderings of the respective spaces. Thus, for each $i = 1, 2$, we write $x \leqq_{K_i} y$ and $y \geqq_{K_i} x$ if and only if $\{x, y\} \subset E_i$ and $y - x \in K_i$.

Theorem 3 **([12, Theorem 49.4]).** *Let the cone K_2 be normal and generating. Furthermore, let $B_k : E_1 \to E_2$, $k = 1, 2$, be additive and homogeneous operators such that B_1^{-1} and $(B_1 + B_2)^{-1}$ exist and possess the properties*

$$B_1^{-1}(K_2) \subset K_1, \tag{19}$$

$$(B_1 + B_2)^{-1}(K_2) \subset K_1 \tag{20}$$

and, furthermore, let the order relation

$$B_1(x - y) \leqq_{K_2} Fx - Fy \leqq_{K_2} B_2(x - y) \tag{21}$$

be satisfied for any pair $(x, y) \in E_1^2$ such that $x \geqq_{K_1} y$.

Then Eq. (18) has a unique solution for an arbitrary z from E_2.

Let us recall two definitions (see, e.g., [11, 12]).

Definition 3. A cone $K_2 \subset E_2$ is called *normal* if there exists a constant $\gamma \in (0, +\infty)$ such that $\|x\|_{E_2} \leq \gamma \|y\|_{E_2}$ for arbitrary $\{x, y\} \subset E_2$ with the property $0 \leqq_{K_2} x \leqq_{K_2} y$.

Definition 4. A cone K_1 is called *generating* in E_1 if every element $u \in E_1$ can be represented in the form $u = u_1 - u_2$, where $\{u_1, u_2\} \subset K_1$.

4.1 Lemmas

We need some technical lemmas.

Lemma 1. *The following propositions are true:*

1. *The set $\mathscr{W}^m_{(0)}$ is a cone in the space $\mathscr{W}^m$.*
2. *The set $\mathscr{W}^m_{(m)}$ is a normal and generating cone in the space $\mathscr{W}^m$.*

Proof. Let us proof assertion 1. If $\{u_1, u_2\} \subset \mathscr{W}^m_{(m)}$ and $\{\lambda_1, \lambda_2\} \subset [0, +\infty)$, then, obviously, $\lambda_1 u_1 + \lambda_2 u_2$ lies in $\mathscr{W}^m_{(m)}$ as well. Suppose that $u \in \mathscr{W}^m_{(m)}$ and $-u \in \mathscr{W}^m_{(m)}$ simultaneously. Taking into account the definition of $\mathscr{W}^m_{(m)}$, we have $u^{(m)} \equiv 0$ and, moreover, $u(a) = 0, \ldots, u^{(m-1)}(a) = 0$, whence it is obvious that $u \equiv 0$. Thus, $\mathscr{W}^m_{(m)}$ is a cone in $\mathscr{W}^m$.

Let us proof assertion 2. In order to check that the cone $\mathscr{W}^m_{(m)}$ is normal, it is sufficient to show that every set of the form

$$\Big\{x \in \mathscr{W}^m : \{x-u, v-x\} \subset \mathscr{W}^m_{(m)}\}, u, v \in W^m, \max\{\|u\|_{\mathscr{W}^m}, \|v\|_{\mathscr{W}^m}\} \leq 1\Big\}, \tag{22}$$

is bounded with respect to the norm $\|\cdot\|_{\mathscr{W}^m}$ (see (6)). Indeed, if an arbitrary x belongs to set (22), then for a.e. $t \in [a,b]$

$$u^{(m)}(t) \leq x^{(m)}(t) \leq v^{(m)}(t), \quad 0 \leq u^{(j)}(a) \leq x^{(j)}(a) \leq v^{(j)}(a),\ 0 \leq j \leq m-1$$

componentwise. Therefore,

$$\|x\|_{\mathscr{W}^m} = \int_a^b \|x^{(m)}(s)\|\,ds + \sum_{i=0}^{m-1} \|x^{(i)}(a)\| \leq \|u\|_{\mathscr{W}^m} + \|v\|_{\mathscr{W}^m} \leq 2,$$

which, in view of the arbitrariness of x, implies that set (22) is bounded.

Finally, let us check, that the cone $\mathscr{W}^m_{(m)}$ is generating cone in the space $\mathscr{W}^m$. To proof that, it is sufficient to show that every element x of $\mathscr{W}^m$ admits a majorant in $\mathscr{W}^m_{(m)}$. Let $x \in \mathscr{W}^m$ be arbitrary. Then x has the form

$$x(t) = \int_a^t \left(\dots \int_a^\alpha X(s)ds \dots\right) d\eta + \sum_{i=0}^{m-1} \frac{(t-a)^{m-i}}{(m-i)!} x^{(m-i)}(a), \qquad t \in [a,b], \tag{23}$$

where $X \in L_1$, $X = x^{(m)}$. Equality (23) implies that, componentwise,

$$x^{(m)}(t) \leq u^{(m)}(t), \qquad t \in [a,b],$$

where for $t \in [a,b]$

$$u(t) = \left(\int_a^t \left(\dots \int_a^\alpha |X_j(s)|ds \dots\right) d\eta + \sum_{i=0}^{m-1} \frac{(t-a)^{m-i}}{(m-i)!} |x_j^{(m-i)}(a)|, \right)_{j=0}^{n}. \tag{24}$$

It is obvious from (24) that $u(a) \geq 0, \dots, u^{(m-1)}(a) \geq 0$, and $u^{(m)}$ is non-negative and, therefore, u is an element of $\mathscr{W}^m_{(m)}$. This, due to the arbitrariness of x, proves that $\mathscr{W}^m_{(m)}$ is generating. The proof is finished.

Let us define a linear operator $V_{l,\overline{h_0,h_{m-1}}} : \mathscr{W}^m \to \mathscr{W}^m$ by putting

$$(V_{l,\overline{h_0,h_{m-1}}}u)(t) := \\ u(t) - \int_a^t \left(\dots \int_a^\alpha (lu)(s)\,ds \dots\right) d\eta - \sum_{i=1}^{m} \frac{(t-a)^{m-i}}{(m-i)!} h_{m-ik}(u) \tag{25}$$

for all $u \in \mathscr{W}^m$.

The next lemmas are true.

Lemma 2. *Function u from the space $\mathscr{W}^m$ is a solution of the equation*

$$(V_{l,\overline{h_0,h_{m-1}}}u)(t) = \int_a^t \left(\dots \int_a^\alpha q(s)ds \dots \right) d\eta, \qquad t \in [a,b],$$

where $q \in \mathscr{L}_1$, if and only if it is a solution of the non-local boundary value problem (3), (4).

The next lemma states the relation between the property described in Definition 1 and the positive invertibility of operator (25).

Lemma 3. *Let $l = (l_k)_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$ is linear operator such that*

$$l \in \mathscr{S}_{\overline{h_0,h_{m-1}}}, \tag{26}$$

then the linear operator $V_{l,\overline{h_0,h_{m-1}}}$: $\mathscr{W}^m \to \mathscr{W}^m$ given by formula (25) *is invertible and, moreover, its inverse $V^{-1}_{l,\overline{h_0,h_{m-1}}}$ is satisfies the inclusion*

$$V^{-1}_{l,\overline{h_0,h_{m-1}}}(\mathscr{W}^m_{(m)}) \subset \mathscr{W}^m_{(0)}. \tag{27}$$

Proof. Suppose that mapping l belongs to the set $\mathscr{S}_{\overline{h_0,h_{m-1}}}$. Given an arbitrary function $y = (y_k)_{k=1}^n \in \mathscr{W}^m$, consider the equation

$$V_{l,\overline{h_0,h_{m-1}}}u = y. \tag{28}$$

Since $y \in \mathscr{W}^m$, then $y^{(m)} \in \mathscr{L}_1$ and

$$y(t) - \sum_{i=0}^{m-1} y^{(m-i)}(a) = \int_a^t \left(\dots \int_a^\alpha y^{(m)}(s)ds \dots \right) d\eta.$$

According to (26), there exists a unique function $u \in \mathscr{W}^m$ such that

$$u^{(m)}(t) = (lu)(t) + y^{(m)}(t), \qquad t \in [a,b],$$
$$u^{(m-i)}(a) = h_{m-i}(u) + y^{(m-i)}(a), \quad i = \overline{1,m}.$$

By Lemma 2, it follows that u is a unique solution of Eq. (28). Due to the arbitrariness of $y \in \mathscr{W}^m$, it follows that $V^{-1}_{l,\overline{h_0,h_{m-1}}}$ exists and, hence, $u = V^{-1}_{l,\overline{h_0,h_{m-1}}}y$.

Inclusion (26) also guarantees that if the functions y_k, $k = \overline{1,n}$, are such that

$$y_k^{(m)}(t) \geq 0,\ y^{(m-1)}(a) \geq 0, \dots,\ y'(a) \geq 0,\ y(a) \geq 0, \tag{29}$$

then the components of u are non-negative and, therefore, $V^{-1}_{l,\overline{h_0,h_{m-1}}}y \in \mathscr{W}^m_{(0)}$. However, relations (29) mean that $y \in \mathscr{W}^m_{(m)}$ (see Notation 4). Since y is arbitrary, we thus arrive at the required inclusion (27).

Lemma 4. *The identity*

$$V_{p,\overline{h_0,h_{m-1}}} + V_{\xi,\overline{h_0,h_{m-1}}} = 2V_{\frac{1}{2}(p+\xi),\overline{\frac{1}{2}(h_0+r_0),\frac{1}{2}(h_{m-1},h_{m-1})}}. \tag{30}$$

holds for arbitrary linear operators $\{p, \xi\} : \mathscr{W}^m \to \mathscr{L}_1$, $i = 1, 2$,

Proof. Equality (30) is obtained immediately from relation (25).

Remark 1. A linear operator $l = (l_k)_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$ belongs to the set $\mathscr{S}_{\overline{h_0,h_{m-1}}}$, if problem

$$u(a) = h(a) \tag{31}$$

for the system

$$\begin{aligned} u_k'(t) = {} & \underbrace{\int_a^t \dots \int_a^\alpha}_{m-1} (l_k u)(s) \underbrace{ds \dots d\eta}_{m-1} + \sum_{i=1}^{m-1} \frac{(t-a)^{m-i}}{(m-i)!} h_{m-ik}(u) \\ & + \underbrace{\int_a^t \dots \int_a^\alpha}_{m-1} q_k(s) \underbrace{ds \dots d\eta}_{m-1}, \quad t \in [a, b], \ k = 1, 2, \dots, n, \end{aligned} \tag{32}$$

has a unique solution $u = (u_k)_{k=1}^n$ for any $\{q_k \mid k = \overline{1, n}\} \subset L_1$ and, moreover, the solution of (32), (31) possesses property (5) if q_k, $k = \overline{1, n}$, are non-negative almost everywhere on $[a, b]$.

A number of results related to the solvability of the linear and non-linear boundary-value problem (32), (31) (and therefore, by virtue of Remark 1, to properties of the set $\mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}}$) can be found, for example, in [2, 3, 6, 7, 13–16].

5 Proof of the Theorem 1

Proof. Let us take $E_1 = E_2 = \mathscr{W}^m$ and define a mapping $F : \mathscr{W}^m \to \mathscr{W}^m$ by setting

$$(Fu)(t) := (V_{f,\overline{\varphi_0,\varphi_{m-1}}}u)(t), \quad t \in [a, b], \tag{33}$$

for any u from $\mathscr{W}^m$, where $V_{f,\overline{\varphi_0,\varphi_{m-1}}}$ is given by (25). Then Eq. (33) takes form (18) with

$$z(t) := \int_a^t \left(\dots \int_a^\alpha q(s) ds \dots \right) d\eta, \qquad t \in [a, b].$$

Consider problem (31), (32). It is clear (see Remark 1) that an absolutely continuous vector function $u = (u_k)_{k=1}^n : [a, b] \to \mathbb{R}^n$ is a solution of (31), (32) if, and only if it satisfies the equation

$$V_{f,\overline{\varphi_0,\varphi_{m-1}}}u = z.$$

Assumption (9) means that the estimate

$$-p_{1k}(u-v)(t) \le -(f_k u)(t) + (f_k v)(t) \le -p_{2k}(u-v)(t), \quad t \in [a,b],$$

is true for any u and v with property (8) and all $k = \overline{1,n}$. The relation

$$u_k^{(m)}(t) - v_k^{(m)}(t) - p_{1k}(u-v)(t) \le u_k^{(m)}(t) - v_k^{(m)}(t) - (f_k u)(t) - (f_k v)(t) \\ \le u_k^{(m)}(t) - v_k^{(m)}(t) - p_{2k}(u-v)(t), \tag{34}$$

hold for almost all t from $[a,b]$.

Let us specify the linear mappings $B_{ik} : \mathscr{W}^m \to \mathscr{W}^m$, $i = 1, 2$, $k = \overline{1,n}$, by the next way

$$(B_{1k}u)(t) := V_{p,\overline{h_0,h_{m-1}}}, \qquad t \in [a,b], \tag{35}$$

$$(B_{2k}u)(t) := V_{\xi,\overline{r_0,r_{m-1}}}, \qquad t \in [a,b], \tag{36}$$

where $\{u, v\} \in \mathscr{W}^m$ have the properties (8). Then m-times integrating (34) and taking property (2) and notation (35), (36) into account, we have

$$B_{1k}(u-v)(t) \\ \le u(t) - \int_a^t \left(\dots \int_a^\alpha (f_k u)(s)ds \dots \right) d\eta - \sum_{i=1}^m \frac{(t-a)^{m-i}}{(m-i)!}\varphi_{m-ik}(u) \\ - (v(t) - \int_a^t \left(\dots \int_a^\alpha (f_k v)(s)ds \dots \right) d\eta - \sum_{i=1}^m \frac{(t-a)^{m-i}}{(m-i)!}\varphi_{m-ik}(v)) \\ \le B_{2k}(u-v)(t), \qquad t \in [a,b], \quad k = \overline{1,n} \tag{37}$$

for any $u = (u_k)_{k=1}^n$ and $v = (v_k)_{k=1}^n$ with properties (8).

In view of the mapping $V_{f,\overline{\varphi_0,\varphi_{m-1}}}$ (see, formulae (25)) and the sets

$$\mathscr{W}^m_{(0)} \text{ and } \mathscr{W}^m_{(m)}$$

(see 5 and 6 in Notation) we have that estimates (34) and (37) ensure the validity of the inclusion

$$B_1(u-v) \le_{\mathscr{W}^m} V_{f,\overline{\varphi_0,\varphi_{m-1}}}u - V_{f,\overline{\varphi_0,\varphi_{m-1}}}v \le_{\mathscr{W}^m} B_2(u-v)$$

for any function u and v with properties (8) from $\mathscr{W}^m$.

Now we determine K_1 and K_2 by the formulae

$$K_1 := \mathscr{W}^m_{(0)}, \qquad K_2 := \mathscr{W}^m_{(m)}. \tag{38}$$

By Lemma 1, the set K_1 forms a cone in the normed space $\mathscr{W}^m$, whereas K_2 is a normal and generating cone in the Banach space $\mathscr{W}^m$.

From Lemma 4 follows, that identity (30) is fulfilled and, therefore,

$$B_1 + B_2 = 2V_{\frac{1}{2}(p_1+p_2),\overline{\frac{1}{2}(h_0+r_0),\frac{1}{2}(h_{m-1}+r_{m-1})}}. \tag{39}$$

Taking into account (7), Lemma 3 guarantees the invertibility of the operators $V_{p_1,\overline{h_0,h_{m-1}}}$ and $V_{\frac{1}{2}(p_1+p_2),\overline{\frac{1}{2}(h_0+r_0),\frac{1}{2}(h_{m-1}+r_{m-1})}}$. So, we have that $B_1^{-1} = V^{-1}_{p_1,\overline{h_0,h_{m-1}}}$ and by (39), the relation

$$(B_1 + B_2)^{-1} = \frac{1}{2}V^{-1}_{\frac{1}{2}(p_1+p_2),\overline{\frac{1}{2}(h_0+r_0),\frac{1}{2}(h_{m-1}+r_{m-1})}}$$

is true. Lemma 3 also ensures the positivity of the inverse operators in the sense that

$$V^{-1}_{p_1,\overline{h_0,h_{m-1}}}(\mathscr{W}^m_{(m)}) \subset \mathscr{W}^m_{(0)},$$
$$V^{-1}_{\frac{1}{2}(p_1+p_2),\overline{\frac{1}{2}(h_0+r_0),\frac{1}{2}(h_{m-1}+r_{m-1})}}(\mathscr{W}^m_{(m)}) \subset \mathscr{W}^m_{(0)}$$

and, hence, inclusions (19), (20) hold.

Finally, in view of assumption (9), we see that relation (21) holds with F, B_1, and B_2 given by (33), (35), (36) with respect to the cones K_1 and K_2 defined by (38).

Applying Theorem 3, we establish the unique solvability of the boundary value problem (32), (31) for arbitrary $q \in \mathscr{L}_1$. Taking Remark 1 into account, we complete the proof of Theorem 1.

6 On Non-local Problems for Higher-Order Linear Functional Differential Equations

Assume that operator f in Eq. (1) and functionals φ_i, $i = \overline{1, m-1}$ in (2) are linear then the next theorem is true.

Theorem 4. *Assume that there exist some linear operators $p = (p_k)_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$, $\xi = (\xi_k)_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$, which satisfy inclusions*

$$p \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}}, \qquad \frac{1}{2}(p+\xi) \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}}, \tag{40}$$

such that inequalities

$$(\xi_k\omega)(t) \le (f_k\omega)(t) \le (p_k\omega)(t), \quad t \in [a,b], \quad k = \overline{1,n},$$

hold for any absolutely continuous vector-function $\omega : [a,b] \to \mathbb{R}^n$ from $\mathscr{W}^m_{(0)}$.

Then the non-local linear boundary-value problem (1), (2) *has a unique solution for an arbitrary function* $q = (q_k)_{k=1}^n \in \mathscr{L}_1$.

Proof. It is easy to see that Theorem 4 is a simple case of the Theorem 1 with

$$u - v = \omega \in \mathscr{W}_{(0)}^m, \quad \varphi_i = h_i = r_i : \mathscr{W}^m \to \mathbb{R}^n, \quad i = \overline{1, m-1}. \tag{41}$$

We get the next corollary.

Corollary 2. *Suppose that for arbitrary absolutely continuous vector-function* $\omega = (\omega_k)_{k=1}^n : [a,b] \to \mathbb{R}^n$ *from* $\mathscr{W}_{(0)}^m$ *the inequality*

$$\big|(f_k\omega)(t) - (l_{1k}\omega)(t)\big| \le (l_{2k}\omega)(t), \quad k = \overline{1,n},$$

holds for some linear operators $l_j = (l_{jk})_{k=1}^n : \mathscr{W}^m \to \mathscr{L}_1$, $j = 1, 2$, *which satisfy inclusions*

$$l_1 + l_2 \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}}, \qquad l_1 \in \mathscr{S}_{(\overline{\varphi_0,\varphi_{m-1}})}. \tag{42}$$

Then the non-local linear boundary-value problem (1), (2) *is uniquely solvable for an arbitrary* $q \in \mathscr{L}_1$.

Proof. Obviously, that Corollary 2 is a simple case of the Theorem 2 with ω, φ_i defined by (41).

7 Optimality of Conditions

Note that assumption (7), (12), (16), (40), (42) we can not replaced by their weaker versions. For example, in Theorem 2 inclusion (42) can not be replaced by the condition

$$(1-\varepsilon)\big(l_1 + l_2\big) \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}}, \quad l_1 \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}} \tag{43}$$

nor by the condition

$$l_1 + l_2 \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}}, \quad (1-\varepsilon)l_1 \in \mathscr{S}_{\overline{\varphi_0,\varphi_{m-1}}} \tag{44}$$

where ε is an arbitrarily small positive number.

To ascertain this we consider the next example.

Example 1. Let us consider the simple linear problem

$$u_1(a) = 0, \quad u_2(a) = 0, \quad u_1'(a) = 0, \quad u_2'(a) = 0, \quad u_1''(a) = 0, \quad u_2''(a) = 0 \tag{45}$$

for the linear functional differential system

$$u_1'''(t) = \frac{1}{(\tau - a)^3}\big(5u_1(\tau) - u_2(\tau)\big) + q_1, \quad t \in [a, b], \tag{46}$$

$$u_2'''(t) = -\frac{1}{(\tau - a)^3}\big(u_1(\tau) - 5u_2(\tau)\big) + q_2, \quad t \in [a, b], \tag{47}$$

where τ is a given point from $(a, b]$, $q_i \in \mathscr{L}_1$, $i = 1, 2$, and fix some $\varepsilon \in [0, 1)$.

It is clear that (45)–(47) is a particular case of problem (1)–(2), where $m = 3$, $n = 2$, operator f is defined by

$$(f_k u)(t) := \frac{(-1)^{k+1}}{(\tau - a)^3}\Big(\big(2k(1 + (-1)^{k+1}) + 1\big)u_1(\tau) - \big(2k + (-1)^k\big)u_2(\tau)\Big),$$

for all $t \in [a, b]$, $k = 1, 2$, and $\varphi_{ik} = 0$, $i = 0, 1, 2$, $k = 1, 2$.

Let us fix some $\varepsilon \in [0, 1)$. It is easy to see that problem (45)–(47) has the family of solutions

$$u_k(t) = (-1)^{k+1}\lambda(t - a)^3, \quad t \in [a, b],\ k = 1, 2,$$

where $\lambda \in \mathbb{R}$ is arbitrary. However, condition (43) in this case is satisfied for all $\varepsilon \in (0, 1)$ with

$$l_1 := 0, \quad l_2 u := \frac{1}{(\tau - a)^2}\begin{pmatrix} 5u_1(\tau) + u_2(\tau) \\ u_1(\tau) + 5u_2(\tau) \end{pmatrix},$$

because initial value problem (45) for the system

$$u_1'''(t) = \frac{(1 - \varepsilon)}{(\tau - a)^2}\Big(5u_1(\tau) + u_2(\tau)\Big) + q_1, \quad t \in [a, b],$$

$$u_2'''(t) = \frac{(1 - \varepsilon)}{(\tau - a)^2}\Big(u_1(\tau) + 5u_2(\tau)\Big) + q_2, \quad t \in [a, b],$$

has only trivial solution.

In a similar way, one can specify an example showing the optimality of condition (44).

Acknowledgements This research was supported in part by the Slovak Grant Agency VEGA-SAV No. 2/0127/20.

References

1. Azbelev, N., Maksimov, V., Rakhmatullina, L.: Introduction to the Theory of Linear Functional Differential Equations. World Federation Publishers Company, Atlanta (1995)
2. Afonso, S.M., Ronto, A.: Measure functional differential equations in the space of functions of bounded variation. Abstr. Appl. Anal. **582161** (2013). https://doi.org/10.1155/2013/582161
3. Baculikova, B., Dzurina, J.: On functional inequalities and their applications in the oscillation theory. Appl. Math. Comput. **226**, 266–273 (2014). https://doi.org/10.1016/j.amc.2013.10.057
4. Bravyi, E.I.: Solvability of the Cauchy problem for higher-order linear functional differential equations. Diff. Equat. **48**(4), 465–476 (2012). https://doi.org/10.1134/S0012266112040015
5. Bravyi, E.I.: On the best constants in the solvability conditions for the periodic boundary value problem for higher-order functional differential equations. Diff. Equat. **48**(6), 779–786 (2012). https://doi.org/10.1134/S001226611206002X
6. Dilnaya, N., Ronto, A.: Multistage iterations and solvability of linear Cauchy problems. Miskolc Math. Notes **4**(2), 89–102 (2003). https://doi.org/10.18514/MMN.2003.81
7. Dilna, N., Ronto, A.: Unique solvability of a non-linear non-local boundary-value problem for systems of non-linear functional differential equations. Math. Slovaca **60**(3), 327–338 (2010). https://doi.org/10.2478/s12175-010-0015-9
8. Dutkiewicz, A.: On the existence of solutions of ordinary differential equations in Banach spaces. Math. Slovaca **65**(3), 573–582 (2015). https://doi.org/10.1515/ms-2015-0041
9. Fečkan, M.: Bifurcation and Chaos in Discontinuous and Continuous Systems. Nonlinear Physical Science. Springer, Heidelberg (2011)
10. Kiguradze, I., Sokhadze, Z.: On nonlinear boundary value problems for higher order functional differential equations. Georgian Math. J. **23**(4), 537–550 (2016). https://doi.org/10.1515/gmj-2016-0039
11. Krasnoselskii, M.A.: Positive Solutions of Operator Equations. Wolters-Noordhoff Scientific Publications, Groningen (1964)
12. Krasnoselskii, M.A., Zabreiko, P.P.: Geometrical Methods of Nonlinear Analysis. Springer, Berlin, New York (1984)
13. Opluštil, Z.: On a nonlocal boundary value problem for first order nonlinear functional differential equations. In: Conference: 3rd International Conference on Differential and Difference Equations and Applications (ICDDEA). Springer Proceedings in Mathematics and Statistics, Mil. Acad, Amadora, Portugal, 05–09 June 2017, vol. 230, pp. 359-371 (2018). https://doi.org/10.1007/978-3-319-75647-9_30
14. Ronto, A.N.: Exact solvability conditions for the Cauchy problem for systems of first-order linear functional-differential equations determined by $(\sigma_1, \sigma_2, \ldots, \sigma_n; \tau)$-positive operators. Ukrainian Math. J. **55**(11), 1541–1568 (2003)
15. Ronto, A.N., Dilna, N.Z.: Unique solvability conditions of the initial value problem for linear differential equations with argument deviations. Nonlinear Oscill. **9**(4), 535–547 (2006). https://doi.org/10.1007/s11072-006-0059-5
16. Ronto, A., Pylypenko, V., Dilna, N.: On the unique solvability of a non-local boundary value problem for linear functional differential equations. Math. Modell. Anal. **13**(2), 241–250 (2008). https://doi.org/10.3846/1392-6292.2008.13.241-250

On the Discrete Fourier Transform Eigenvectors and Spontaneous Symmetry Breaking

Mesuma K. Atakishiyeva, Natig M. Atakishiyev, and Juan Loreto-Hernández

Abstract The present work aims to give a detailed discussion of a recently introduced difference analogue of quantum number operator in terms of the raising and lowering difference operators, that governs eigenvectors of the N-dimensional discrete (finite) Fourier transform (DFT). In particular, we argue that the aforementioned discrete number operator $\mathscr{N}^{(N)}$ has distinct eigenvalues only if it is associated with the DFT's based on grids $\{x_0, x_1, ..., x_{N-1}\}, x_k = \sqrt{2\pi/N}\,k$, with N odd. This means that in the cases of the DFT's on grids $\{x_0, x_1, ..., x_{N-1}\}$ with N even the discrete reflection symmetry in the space of eigenvectors of the discrete number operator $\mathscr{N}^{(N)}$ is spontaneously broken. This essential distinction between even and odd dimensions is intimately related with the algebraic properties of the above DFT raising and lowering difference operators and consistent with the well-known formula for the multiplicities of the eigenvalues, associated with the N-dimensional discrete Fourier transform.

1 Introduction

We choose to set out by recalling first some mathematical aspects of the classical Fourier integral transform (FIT)

M. K. Atakishiyeva
Universidad Autónoma del Estado de Morelos, Centro de Investigación en Ciencias,
62250 Cuernavaca, Morelos, Mexico
e-mail: mesuma@uaem.mx

N. M. Atakishiyev (✉)
Universidad Nacional Autónoma de México, Instituto de Matemáticas,
Unidad Cuernavaca, 62210 Cuernavaca, Morelos, Mexico
e-mail: natig@im.unam.mx

J. Loreto-Hernández
Universidad Autónoma de San Luis Potosí, Facultad de Ciencias,
78290 San Luis Potosí, SLP, Mexico
e-mail: juan@fc.uaslp.mx

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_41

$$(\mathscr{F} f)(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\,\mathrm{i}\,x\,y} f(y)\, dy = g(x) \tag{1.1}$$

and its finite analogue, discrete Fourier transform (DFT) $\Phi^{(N)}$. It is well known that the eigenfunctions of the FIT operator (1.1), associated with the eigenvalues i^n, are explicitly given as

$$\psi_n(x) := H_n(x) \exp\left(-x^2/2\right), \tag{1.2}$$

where $H_n(x)$ are the Hermite polynomials. Recall that the functions $\psi_n(x)$ are usually referred to as *Hermite functions* in the mathematical literature, whereas in quantum mechanics they emerge as eigenfunctions of the Hamiltonian $\mathbf{H}$ for the linear harmonic oscillator, which is a self-adjoint differential operator of the second order (see, for example, [1]),

$$\mathbf{H} = \frac{1}{2}\left(x^2 - \frac{d^2}{dx^2}\right). \tag{1.3}$$

The functions $\psi_n(x)$ represent an important explicit example of an orthogonal and complete system in the Hilbert space $L^2(\mathbb{R}, dx)$ of square-integrable functions on the full real line $x \in \mathbb{R}$. Observe also that since the fourth power of the FIT operator $\mathscr{F}$ is the unit operator, the only four distinct eigenvalues among i^ns are ± 1 and $\pm\mathrm{i}$.

As for the discrete Fourier transform $\Phi^{(N)}$, based on N points, it is represented by the $N \times N$ unitary symmetric matrix $\Phi^{(N)} = \| \Phi^{(N)}_{m,n} \|$ (frequently referred to as Schur's matrix) with elements

$$\Phi^{(N)}_{m,n} = \frac{1}{\sqrt{N}} \exp\left(\frac{2\pi \mathrm{i}}{N}\, m\, n\right) \equiv \frac{1}{\sqrt{N}}\, q^{mn}, \tag{1.4}$$

where $q := \exp(2\pi\mathrm{i}/N)$ and $m,\ n \in \{0, 1, \ldots, N-1\}$ [2–6]. Given a complex valued vector $\mathbf{y}$ with components $\{y_k\}_{k=0}^{N-1}$, one can compute another vector $\mathbf{z}$ with components

$$z_m = \sum_{n=0}^{N-1} \Phi^{(N)}_{m,n}\, y_n, \tag{1.5}$$

referred to as *the discrete Fourier transform* of the vector $\mathbf{y}$. Those vectors $\mathbf{f}^{(k)}$, which are solutions of the standard equations

$$\sum_{n=0}^{N-1} \Phi^{(N)}_{m,n} f_n^{(k)} = \lambda_k f_m^{(k)}, \qquad k \in \{0, 1, \ldots, N-1\}, \tag{1.6}$$

then represent eigenvectors of the DFT operator $\Phi^{(N)}$, associated with the eigenvalues λ_k. Since the fourth power of $\Phi^{(N)}$ is the unit matrix, the only four distinct eigenvalues among λ_k's are $\pm\, 1$ and $\pm\, \mathrm{i}$, the same as in the case of the FIT operator $\mathscr{F}$.

The purpose of this presentation is to provide a detailed account of a difference analogue of the differential operator $\mathbf{H}$, that commutes with the DFT operator $\Phi^{(N)}$. The ability to solve a difference equation for this difference analogue of $\mathbf{H}$, then enables one to find an analytical form of the set of eigenvectors for the DFT operator $\Phi^{(N)}$. The key to our approach thus lies in the recognition that the eigenvectors of the DFT operator $\Phi^{(N)}$ can be constructed in complete analogy with the continuous case of the FIT operator $\mathscr{F}$ (see [7–9] for details of this approach to deriving the DFT eigenvectors). The 'value added' in moving from the continuous FIT case to its discrete counterpart DFT is the discovery that the discrete number operator $\mathscr{N}^{(N)}$, that governs eigenvectors of the N-dimensional DFT, has distinct eigenvalues only if it is associated with the DFT's based on grids $\{0, 1, ..., N-1\}$ with odd $N = 2L + 1$, whereas in the cases of the DFT's on grids $\{0, 1, ..., N-1\}$ with even $N = 2L$ the discrete reflection symmetry in the space of eigenvectors of the discrete number operator $\mathscr{N}^{(N)}$ is spontaneously broken. This essential distinction between even and odd dimensions is shown to be consistent with the well-known formula for the multiplicities of the eigenvalues, associated with the N-dimensional DFT [10].

In Sect. 2 we discuss how to construct a difference equation for the DFT eigenvectors in terms of the lowering and raising difference operators, which are defined by the standard intertwining relations with the DFT operator $\Phi^{(N)}$. Section 3 is devoted to the evaluation of the rank of those lowering and raising operators. Section 4 discusses the consistency of our approach with the well-known formula for the multiplicities of the eigenvalues, associated with the N-dimensional DFT. In Section 5, we derive an explicit form of the lowest odd-dimensional DFT eigenvector in terms of finite continuous fractions. Finally, Sect. 6 briefly outlines some further research directions of interest.

2 Difference Equation for the DFT Eigenvectors

It is a remarkable fact that an appropriate difference equation for the eigenvectors $\mathbf{f}^{(n)}$ of the DFT, associated with the eigenvalues $\lambda_n = \mathrm{i}^n$, can be constructed in complete analogy with the continuous case when the functions $\psi_n(x)$, defined as in (1.2), satisfy the second order differential equation

$$\mathbf{H}\,\psi_n(x) = (n + 1/2)\,\psi_n(x). \tag{2.1}$$

To bring out this fact we recall first that the differential operator $\mathbf{H}$ in (2.1) can be factorized in terms of the lowering $\mathbf{a}$ and raising $\mathbf{a}^{\dagger}$ first order differential operators as

$$\mathbf{H} = \mathbf{a}^{\dagger}\,\mathbf{a} + \frac{1}{2}\,\mathbf{I}, \tag{2.2}$$

where **I** is the identity operator and

$$\mathbf{a} = \frac{1}{\sqrt{2}}\left(x + \frac{d}{dx}\right), \qquad \mathbf{a}^{\dagger} = \frac{1}{\sqrt{2}}\left(x - \frac{d}{dx}\right). \tag{2.3}$$

The operators **a** and $\mathbf{a}^{\dagger}$ obey the standard Heisenberg commutation relation

$$\left[\mathbf{a},\, \mathbf{a}^{\dagger}\right] := \mathbf{a}\,\mathbf{a}^{\dagger} \,-\, \mathbf{a}^{\dagger}\,\mathbf{a} \,\equiv \left[\frac{d}{dx},\, x\right] = \mathbf{I} \tag{2.4}$$

and the intertwining relations

$$\mathbf{a}\,\mathscr{F} = \mathrm{i}\,\mathscr{F}\,\mathbf{a}, \qquad \mathbf{a}^{\dagger}\,\mathscr{F} = -\,\mathrm{i}\,\mathscr{F}\mathbf{a}^{\dagger}, \tag{2.5}$$

with the FIT operator $\mathscr{F}$. Recall also that very often it is more convenient to work directly with the quantum number operator **N**, defined as

$$\mathbf{N} := \mathbf{H} \,-\, \frac{1}{2}\mathbf{I} \,=\, \mathbf{a}^{\dagger}\,\mathbf{a}. \tag{2.6}$$

Then from (2.5) it follows at once that the differential operator **H**, as well as the number operator **N**, do commute with the FIT operator $\mathscr{F}$. This is why the FIT operator $\mathscr{F}$ and the differential operator **H** (or, equivalently, the number operator **N**) have the same set of the eigenfunctions $\psi_n(x)$, which can be now written in a compact form as

$$\psi_n(x) \,=\, \frac{1}{\sqrt{n!}}\,\left(\mathbf{a}^{\dagger}\right)^n\,\psi_0(x). \tag{2.7}$$

Recently it was suggested in [7] to construct a discrete number operator

$$\mathscr{N}^{(N)} := \mathbf{b}_N^{\mathsf{T}}\,\mathbf{b}_N \tag{2.8}$$

in terms of the difference raising and lowering operators $\mathbf{b}_N^{\mathsf{T}}$ and $\mathbf{b}_N$, which are defined by the standard intertwining relations

$$\mathbf{b}_N\,\Phi^{(N)} = \mathrm{i}\,\Phi^{(N)}\,\mathbf{b}_N, \qquad \mathbf{b}_N^{\mathsf{T}}\,\Phi^{(N)} = -\,\mathrm{i}\,\Phi^{(N)}\,\mathbf{b}_N^{\mathsf{T}}, \qquad N \geq 3, \tag{2.9}$$

with the DFT operator $\Phi^{(N)}$. Then from (2.9) it follows at once that this discrete number operator $\mathscr{N}^{(N)}$ does commute with the DFT operator $\Phi^{(N)}$ and can be interpreted as a difference analogue of the quantum number operator **N**, associated with the continuous IFT. Therefore solutions of a difference equation for eigenvectors of the discrete number operator $\mathscr{N}^{(N)}$ represent the desired set of eigenvectors for the DFT operator $\Phi^{(N)}$ at the same time.

We now turn to the question of finding explicit forms of the difference raising and lowering operators $\mathbf{b}_N^{\mathsf{T}}$ and $\mathbf{b}_N$, defined by the (2.9). Let us introduce first two operators $\mathbf{Q}^{(\pm)}$ with matrix elements $Q_{kl}^{(\pm)} := q^{\pm k}\delta_{kl}$, where $q = e^{2\pi \mathrm{i}/N} \equiv e^{\mathrm{i}\,\theta_N}$, $\theta_N := 2\pi/N$. Then, by the above definition, $\mathbf{Q}^{(\pm)}\,\mathbf{y} = \mathbf{z}^{(\pm)}$, where $z_k^{(\pm)} = \sum_{l=0}^{N-1} Q_{kl}^{(\pm)}\,y_l = \sum_{l=0}^{N-1} q^{\pm k}\delta_{kl}\,y_l = q^{\pm k}\,y_k$. This means that under the action of the operators $\mathbf{Q}^{(\pm)}$ the components y_k of an arbitrary vector $\mathbf{y}$ get multiplied by the factors $q^{\pm k}$, respectively. Consequently, for their linear combinations

$$\mathbf{C} := \frac{1}{2}\Big(\mathbf{Q}^{(+)} + \mathbf{Q}^{(-)}\Big), \qquad \mathbf{S} := \frac{\mathbf{1}}{\mathbf{2i}}\Big(\mathbf{Q}^{(+)} - \mathbf{Q}^{(-)}\Big), \tag{2.10}$$

one obtains that

$$\begin{aligned}\Big(\mathbf{C}\,\mathbf{y}\Big)_k &= \frac{1}{2}\left(q^k + q^{-k}\right) y_k = \cos(k\,\theta_N)\,y_k,\\ \Big(\mathbf{S}\,\mathbf{y}\Big)_k &= \frac{1}{2\mathrm{i}}\left(q^k - q^{-k}\right) y_k = \sin(k\,\theta_N)\,y_k.\end{aligned} \tag{2.11}$$

The next step is to define a pair of the shift operators $\mathbf{T}^{(\pm)}$ with matrix elements $T_{kl}^{(\pm)} := \delta_{k\pm1,\,l}$, where $\delta_{-1,\,l} \equiv \delta_{N-1,\,l}$ and $\delta_{N,\,l} \equiv \delta_{0,\,l}$. Then

$$\Big(\mathbf{T}^{(\pm)}\,\mathbf{y}\Big)_k = \sum_{l=0}^{N-1} T_{kl}^{(\pm)}\,y_l = \sum_{l=0}^{N-1} \delta_{k\pm1,\,l}\,y_l = y_{k\pm1}, \tag{2.12}$$

where $y_{-1} \equiv y_{N-1}$ and $y_N \equiv y_0$.

Lemma 1. *Components of any N-periodic vector* $\mathbf{y}$ *satisfy the following identities:*

$$\sum_{k=0}^{N-1} q^{j(k\pm1)}\,y_k = \sum_{k=0}^{N-1} q^{jk}\,y_{k\mp1}, \qquad j = 0, 1, 2, ..., N-1. \tag{2.13}$$

Proof. Begin with the left side of the first identity in (2.13),

$$\sum_{k=0}^{N-1} q^{j\,(k+1)}\,y_k = \sum_{m=1}^{N} q^{j\,m}\,y_{m-1} = \sum_{m=1}^{N-1} q^{j\,m}\,y_{m-1} + q^{j\,N}\,y_{N-1} = \sum_{m=0}^{N-1} q^{j\,m}\,y_{m-1}, \tag{2.14}$$

in view of the evident relation $q^{Nj} = 1$. The second identity in (2.13) is argued similarly. □

The operators $\mathbf{Q}^{(\pm)}$ and $\mathbf{T}^{(\pm)}$ are actually interconnected through the FFT operator $\Phi^{(N)}$ in the following way.

Proposition 1. *The intertwining relations*

$$\mathbf{Q}^{(\pm)}\,\Phi^{(N)} = \Phi^{(N)}\,\mathbf{T}^{(\mp)}, \qquad \mathbf{T}^{(\pm)}\,\Phi^{(N)} = \Phi^{(N)}\,\mathbf{Q}^{(\pm)}, \tag{2.15}$$

for the operators $\mathbf{Q}^{(\pm)}$ *and* $\mathbf{T}^{(\pm)}$ *with the FFT operator* $\Phi^{(N)}$ *are respectively valid.*

Proof. Two relations on the left in (2.15) follow directly from the lemma, if one takes into account the definition of the shift operators $\mathbf{T}^{(\pm)}$. So let us consider now matrix elements of two identities on the right in (2.15):

$$\left(\mathbf{T}^{(\pm)}\,\Phi^{(N)}\right)_{mn} = \sum_{k=0}^{N-1} T^{(\pm)}_{mk}\,\Phi^{(N)}_{kn} = \sum_{k=0}^{N-1} \delta_{m\pm 1,\,k}\,\Phi^{(N)}_{kn} = \Phi^{(N)}_{m\pm 1,\,n}$$
$$= \frac{1}{\sqrt{N}}\,q^{(m\pm 1)\,n} = q^{\pm n}\,\Phi^{(N)}_{m,\,n} = \sum_{l=0}^{N-1} \Phi^{(N)}_{m,\,l}\,q^{\pm l}\,\delta_{l\,n} = \left(\Phi^{(N)}\,\mathbf{Q}^{(\pm)}\right)_{mn}.$$

This completes the proof of the proposition. □

Notice that if a vector $\mathbf{y}^{(n)}$ with the components $\{y^{(n)}_j\}_{j=0}^{N-1}$ is N-periodic, then all m vectors $\mathbf{z}^{(n;\,m)}$, which have, by definition, components $\{z^{(n;\,m)}_j\}_{j=0}^{N-1}{:=}\{q^{mj}\,y^{(n)}_j\}_{j=0}^{N-1}$, m is an arbitrary integer number, are also N-periodic. Therefore, applying each line in (2.15) m times in succession, one arrives at the following corollary to the proposition:

$$q^{\pm mj}\sum_{k=0}^{N-1} \Phi^{(N)}_{j,\,k}\,y^{(n)}_k = \sum_{k=0}^{N-1} \Phi^{(N)}_{j,\,k}\,\mathbf{T}^m_{\mp}\,y^{(n)}_k,$$
$$\mathbf{T}^m_{\pm}\sum_{k=0}^{N-1} \Phi^{(N)}_{j,\,k}\,y^{(n)}_k = \sum_{k=0}^{N-1} \Phi^{(N)}_{j,\,k}\,q^{\pm mk}\,y^{(n)}_k, \tag{2.16}$$

where m is an integer.

Finally, to formulate explicit forms of the operators $\mathbf{b}_N$ and $\mathbf{b}^{\mathsf{T}}_N$, let us consider now two operators $\mathscr{X}$ and $\mathscr{D}$ ($N > 2$):

$$\mathscr{X} := \frac{1}{2\mathrm{i}}\sqrt{\frac{N}{2\pi}}\left(\mathbf{Q}^{(+)} - \mathbf{Q}^{(-)}\right) \equiv \sqrt{\frac{N}{2\pi}}\,\mathbf{S}, \quad \mathscr{D} := \frac{1}{2}\sqrt{\frac{N}{2\pi}}\left(\mathbf{T}^{(+)} - \mathbf{T}^{(-)}\right). \tag{2.17}$$

From the intertwining relations (2.15) it follows then that

$$\mathscr{D}\,\Phi^{(N)} = \mathrm{i}\,\Phi^{(N)}\,\mathscr{X}, \qquad \mathscr{X}\,\Phi^{(N)} = \mathrm{i}\,\Phi^{(N)}\,\mathscr{D}. \tag{2.18}$$

Difference raising and lowering operators $\mathbf{b}_N$ and $\mathbf{b}^{\mathsf{T}}_N$ may be now defined as

$$\mathbf{b}_N := \frac{1}{\sqrt{2}}\,(\mathscr{X} + \mathscr{D}) = \frac{1}{2}\sqrt{\frac{N}{\pi}}\left[\mathbf{S} + \frac{1}{2}\left(\mathbf{T}^{(+)} - \mathbf{T}^{(-)}\right)\right], \tag{2.19a}$$

$$\mathbf{b}^{\mathsf{T}}_N := \frac{1}{\sqrt{2}}\,(\mathscr{X} - \mathscr{D}) = \frac{1}{2}\sqrt{\frac{N}{\pi}}\left[\mathbf{S} - \frac{1}{2}\left(\mathbf{T}^{(+)} - \mathbf{T}^{(-)}\right)\right]. \tag{2.19b}$$

We also display the $N \times N$ matrix form of the operators $\mathbf{b}_N$ and $\mathbf{b}_N^{\mathsf{T}}$, respectively:

$$\mathbf{b}_N = \sqrt{\frac{2\pi}{N}} \begin{bmatrix} 0 & 1 & 0 & 0 \cdots & 0 & 0 & 0 & -1 \\ -1 & s_1 & 1 & 0 \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & s_2 & 1 \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & s_3 \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \cdots & s_{N-4} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \cdots & -1 & s_{N-3} & 1 & 0 \\ 0 & 0 & 0 & 0 \cdots & 0 & -1 & s_{N-2} & 1 \\ 1 & 0 & 0 & 0 \cdots & 0 & 0 & -1 & s_{N-1} \end{bmatrix}, \tag{2.20}$$

where $s_k := 2 \sin \frac{2\pi}{N} k$, $0 \le k \le N-1$, and N is an arbitrary integer, $N \ge 2$. Notice that the raising difference operator $\mathbf{b}_N^{\mathsf{T}}$ is the matrix transpose of the lowering difference operator $\mathbf{b}_N$.

The DFT eigenvectors $\mathbf{f}_n^{(N)}$ are thus governed by the difference equation

$$\mathscr{N}^{(N)} \mathbf{f}_n^{(N)} = \lambda_n^{(N)} \mathbf{f}_n^{(N)}, \tag{2.21}$$

where the discrete number operator $\mathscr{N}^{(N)}$ is expressed in terms of the raising and lowering difference operators $\mathbf{b}_N^{\mathsf{T}}$ and $\mathbf{b}_N$ as $\mathscr{N}^{(N)} = \mathbf{b}_N^{\mathsf{T}}\, \mathbf{b}_N$.

3 More on Raising and Lowering Operators

In this section we evaluate the rank of the N-dimensional difference operators $\mathbf{b}_N$ and $\mathbf{b}_N^{\mathsf{T}}$ and demonstrate that the characteristic equations for those operators have a particular ‘cyclic’ form. To begin with, notice that the $N \times N$ matrix in (2.20), associated with the lowering difference operator $\mathbf{b}_N$, is traceless, because one checks easily that $\sum_{k=1}^{N-1} s_k = 0$. Also, since $s_{N-k} = -s_k$ by definition of the parameters s_k, the diagonal elements of the matrix in (2.20) are of the form

$$\begin{aligned} \{0, s_1, s_2, ..., s_{L-1}, 0, -s_{L-1}, ..., -s_2, -s_1\}, \quad N = 2L, \\ \{0, s_1, s_2, ..., s_L, -s_L, ..., -s_2, -s_1\}, \quad N = 2L+1, \end{aligned} \tag{3.1}$$

for even and odd dimensions N, respectively. Observe also that this matrix $\mathbf{b}_N$ is of ‘almost’ tridiagonal form: it has ± 1 elements in the upper-right and lower-left corners but otherwise is tridiagonal. Since those ± 1 elements can be regarded as cyclic extensions of the subdiagonal and the superdiagonal elements, these type of matrices are referred to as *extended-tridiagonal* matrices [11–13].

The matrix $\mathbf{b}_N$ is noninvertible and its rank is different for the even and odd dimensions N. This can be shown in the following way.

At this point, we recall first that the space spanned by the *rows (columns)* of a matrix A is called the *row (column) space* of A; its dimension is called the *row* *(column) rank*. The rank of a matrix equals the row (column) rank [14]. Also, the process of Gaussian elimination is simplified by working directly with the augmented matrix of the linear system and applying certain operations to its rows (columns). Those row (column) operations correspond to the three types of operation that may be applied to a linear system during Gaussian elimination: they are called *elementary row (column) operations* and they can be applied to any matrix. The elementary row (column) operations are as follows:

(a) interchange rows (columns) i and j;
(b) add c times row (column) j to row (column) i where c is any scalar;
(c) multiply row (column) i by a non-zero scalar c.

From the matrix point of view the essential content of theorem, which describes the possible behaviour of a linear system, is that any matrix can be put in *row echelon form* by application of a suitable finite sequence of elementary row operations [15].

Let us evaluate now the rank of the matrix $\mathbf{b}_N = \sqrt{2\pi/N} \times M_L$ with odd $N = 2L+1$, $L \geq 1$, where the matrix M_L is

$$M_L := \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 \\ -1 & s_1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & s_2 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & s_3 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -1 & s_L & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & -1 & -s_L & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & -1 & -s_3 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & -s_2 & 1 \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & -s_1 \end{bmatrix}. \qquad (3.2)$$

By elementary column operations [15] one may cast the matrix M_L into the form

$$M_L^{'} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \\ s_1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1 \\ -1 & s_2 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & s_3 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -1 & s_L & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 0 & -1 & -s_L & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & -1 & -s_3 & 1 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & -s_2 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & -s_1 & 1 \end{bmatrix}, \qquad (3.3)$$

which is almost in the column echelon form, except for the two elements equal to -1 on the upper right corner of this matrix. But it is not hard to eliminate them by further elementary column operations over the last two columns, and bring $M_L^{'}$ into complete column echelon form. Indeed, let us add to the penultimate column in (3.3) its first column plus the last one, multiplied by the parameter s_1, to get

$$\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0\\
s_1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1\\
-1 & s_2 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & 0\\
0 & -1 & s_3 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0\\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
0 & \cdots & 0 & -1 & s_L & 1 & 0 & \cdots & 0 & 0 & 0\\
0 & \cdots & 0 & 0 & -1 & -s_L & 1 & \cdots & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots\\
0 & \cdots & 0 & 0 & 0 & \cdots & -1 & -s_3 & 1 & 0 & 0\\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & -s_2 & 1 & 0\\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 0 & 1
\end{bmatrix}. \tag{3.4}$$

So the next step is to add to the last column in (3.4) the second column plus the penultimate one, multiplied by the parameter s_2; this results in

$$M_L^{''} = \left[\begin{array}{cc|ccccccccc}
1 & \multicolumn{1}{c}{0} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0\\
s_1 & \multicolumn{1}{c}{1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0\\
\cline{3-11}
-1 & s_2 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & 0\\
0 & -1 & s_3 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1\\
\vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
0 & \cdots & 0 & -1 & s_L & 1 & 0 & \cdots & 0 & 0 & 0\\
0 & \cdots & 0 & 0 & -1 & -s_L & 1 & \cdots & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots\\
0 & \cdots & 0 & 0 & 0 & \cdots & -1 & -s_3 & 1 & 0 & 0\\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & -1 & -s_2 & 1 & s_2\\
0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 0 & 0
\end{array}\right]. \tag{3.5}$$

Let us emphasize that this form $M_L^{''}$ of the initial matrix M_L contains the principal minor $(M_L^{''})_{i,j}$, $3 \leq i, j \leq N$, of order $N - 2 \equiv 2(L - 1) + 1$, which results from the deletion of the first two rows and columns in (3.5), and this minor has the same structure as the matrix $M_{L-1}^{'}$ in (3.3). One therefore may employ the same type of elementary column operations as above in order to move those two elements equal to -1 in the last two columns in (3.5) another two rows down. Repeating those steps $L - 1$ times, one finally arrives at the form, which contains a principal minor of order 3 on its lower right corner; this minor has the same structure as $M_1^{'}$. It is plain that in the latter case

$$M_1^{'} = \begin{bmatrix} 1 & -1 & 0 \\ s_1 & 1 & -1 \\ -1 & -s_1 & 1 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ s_1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \Longrightarrow \begin{bmatrix} 1 & 0 & 0 \\ s_1 & 1 & 0 \\ -1 & -1 & 0 \end{bmatrix}, \tag{3.6}$$

where we added the first column plus the third column, multiplied by s_1, to the second column at the first step, and the second column to the third one at the second step. This means that the matrix M_L is finally reduced to the complete column echelon form of a lower triangular matrix, which has nonzero first $N-1$ diagonal elements and vanishing the last diagonal element and one zero elements on its main diagonal. Hence $\det M_L = 0$, $\operatorname{rank} M_L = N-1$ and the null space of the matrix M_L is one-dimensional.

Turning now to the case of the matrix $\mathbf{b}_N$ with even $N = 2L$, $L \geq 2$, one may likewise employ the same inductive procedure in L as in the odd case $N = 2L+1$. Although the corresponding matrix for even $N = 2L$ has almost the same structure as M_L in the odd case, the diagonal elements of those two matrices are essentially different: as indicated in (3.1), there are two zero diagonal elements in the even case $N = 2L$ and only one zero element in the odd case $N = 2L+1$. One therefore checks easily that all even case matrices can be transformed, by elementary column operations, into the complete echelon form with two zero columns at the end. Thus in the even $N = 2L$ case $\det \mathbf{b}_N = 0$, $\operatorname{rank} \mathbf{b}_N = N-2$ and the null space of the matrix $\mathbf{b}_N$ is two-dimensional.

One important remark must be made at this point, however. Since the discrete quantum number operator $\mathcal{N}^{(N)}$ is defined as $\mathcal{N}^{(N)} = \mathbf{b}_N^{\mathsf{T}}\, \mathbf{b}_N$, every element of the null space of the matrix $\mathbf{b}_N$ evidently represents at the same time the eigenvector of the operator $\mathcal{N}^{(N)}$, associated with the zero eigenvalue as well. Therefore the present result that the dimensions of the null spaces of the matrices $\mathbf{b}_N$ are equal to 1 and 2 for odd and even N's, respectively, already strongly indicates that the spectral properties of the discrete quantum number operator $\mathcal{N}^{(N)}$ essentially differ for even and odd dimensions N. We shall comment more on this heretofore undetected distinction between even and odd dimensional DFT's below, in Sect. 5.

Let us draw attention now to the remarkable 'cyclic' properties of the lowering and raising difference operators $\mathbf{b}_N$ and $\mathbf{b}_N^{\mathsf{T}}$, which can be formulated in the following way. Recall first that the equation which is solved to find eigenvalues of $N \times N$ matrix M is usually interpreted as the equation for finding roots of the characteristic polynomial

$$p_M(\lambda) := det(\lambda I - M) = \lambda^N + c_1\lambda^{N-1} + c_2\lambda^{N-2} + \cdots + c_{N-1}\lambda + c_N, \tag{3.7}$$

where I is the $N \times N$ identity matrix and the coefficient c_k is $(-1)^k$ times the sum of the determinants of all of the principal $k \times k$ minors of M (note that by this definition $c_1 = -\,trace(M)$ and $c_N = (-1)^N detM$). The Cayley–Hamilton theorem then states that a $N \times N$ matrix M is annihilated by its characteristic polynomial (3.7), that is,

$$p_M(M) = M^N + c_1 M^{N-1} + c_2 M^{N-2} + \cdots + c_{N-1} M + c_N = 0. \tag{3.8}$$

It is plain that for a singular traceless matrix M the coefficients c_1 and c_N vanish. It turns out that for the particular traceless matrices of the form $\mathbf{b}_N$ and $\mathbf{b}_N^{\mathsf{T}}$ there are many more zero coefficients in the identity (3.8). The point is that from the defining intertwinning relations (2.9) for the lowering and raising difference operators $\mathbf{b}_N$ and $\mathbf{b}_N^{\mathsf{T}}$ it follows that

$$(\Phi^{(N)})^{\dagger}\,\mathbf{b}_N\,\Phi^{(N)} = \mathrm{i}\,\mathbf{b}_N, \qquad (\Phi^{(N)})^{\dagger}\,\mathbf{b}_N^{\mathsf{T}}\,\Phi^{(N)} = -\,\mathrm{i}\,\mathbf{b}_N^{\mathsf{T}}. \tag{3.9}$$

This means that if, for instance, $\mathbf{b}_N\,\mathbf{f}^{(N)} = \lambda\,\mathbf{f}^{(N)}$ and $\lambda \neq 0$, then $\mathbf{b}_N\,\mathbf{g}^{(N)} = -\mathrm{i}\lambda\,\mathbf{g}^{(N)}$, where the eigenvector $\mathbf{g}^{(N)}$ of the operator $\mathbf{b}_N$ is defined as $\mathbf{g}^{(N)} := (\Phi^{(N)})^{\dagger}\,\mathbf{f}^{(N)}$. Hence, if the operator $\mathbf{b}_N$ has nonzero eigenvalue λ, associated with the eigenvector $\mathbf{f}^{(N)}$, then it has another eigenvalue $-\mathrm{i}\lambda$, associated with the eigenvector $\mathbf{g}^{(N)}$, as well. Moreover, since the DFT operator $\Phi^{(N)}$ is of order 4, each nonzero eigenvalue λ is actually accompanied by the 3 other eigenvalues $\mathrm{i}^k\,\lambda, k = 1, 2, 3$. Since the polynomial $(z-\lambda)(z-\mathrm{i}\lambda)(z+\lambda)(z+\mathrm{i}\lambda) = z^4 - \lambda^4$, one therefore concludes that in the characteristic Eq. (3.8) for the lowering operator $\mathbf{b}_N, N \geq 5$, the only nonzero coefficients are $c_4, c_8, ..., c_{4k}$, where $k := [N/4]$. This characteristic equation can be thus written in the 'cyclic' form as

$$\mathbf{b}_N^N + c_4\mathbf{b}_N^{N-4} + ... + c_{4k}\mathbf{b}_N^{N-4k} = \left(\mathbf{b}_N^4 - \lambda_1^4\right)\left(\mathbf{b}_N^4 - \lambda_2^4\right)...\left(\mathbf{b}_N^4 - \lambda_k^4\right)\mathbf{b}_N^l = 0, \tag{3.10}$$

where $0 \leq l := N - 4k \leq 3$ and $\lambda_1, \lambda_2, ..., \lambda_k$ are some constants.

To close this section, we recall here that the raising difference operator $\mathbf{b}_N^{\mathsf{T}}$ is the matrix transpose of the lowering difference operator $\mathbf{b}_N$; hence the former operator has the same set of properties as the latter one: the vanishing determinant and trace, the same rank, distinct for even and odd dimensions N, and the same characteristic Eq. (3.10).

4 Multiplicities of the DFT Eigenvalues

As a starting point in this section, we remind the reader that there are the two so-called Chebyshev sets of functions

$$\{\sin t, \sin 2t, ..., \sin nt\}, \qquad \{1, \cos t, \cos 2t, ..., \cos nt\}, \tag{4.1}$$

which are defined on intervals $(0, \pi)$ and $[0, \pi]$, respectively. These Chebyshev sets were employed in [2, 3] to give a simple proof of the explicit expressions for the multiplicities $m_k(\mathrm{i}^k),\ 0 \leq k \leq 3$ of the eigenvalues $1, \mathrm{i}, -1, -\mathrm{i}$ of the N-dimensional discrete Fourier transform,

$$m_0(1) = \left[\frac{N}{4}\right] + 1, \qquad m_1(\mathrm{i}) = \left[\frac{N+1}{4}\right],$$
$$m_2(-1) = \left[\frac{N+2}{4}\right], \qquad m_3(-\mathrm{i}) = \left[\frac{N+3}{4}\right] - 1, \tag{4.2}$$

where the symbol $[X]$ stands for the greatest integer lower than X or equal to X. Note that the lowering $\mathbf{b}_N$ and raising $\mathbf{b}_N^{\mathsf{T}}$ difference operators, as well as their product, the discrete number operator $\mathscr{N}^{(N)} = \mathbf{b}_N^{\mathsf{T}}\mathbf{b}_N$, do depend on the set of parameters $\{s_1, s_2, ..., s_{N-1}\}$, which may be regarded as the particular case of the Chebyshev set of smooth functions taken at the distinct points $t_k := 2\pi k/N$, $1 \le k \le (N-1)$. Thus the fact that this particular set of parameters $\{s_1, s_2, ..., s_{N-1}\}$ plays a key role in our study, is not accidental at all. But the real reason for mentioning here the formula (4.2) for multiplicities is the following. The formula (4.2) had been known for a long time before the appearance of the above-mentioned papers [2, 3]. Nevertheless, what seems to remain unnoticed is that this formula clearly points out the dissimilarity between ordering the even-dimensional eigenvectors and odd-dimensional eigenvectors of the DFT operator $\Phi^{(N)}$. This can be argued in the following way.

Imagine that one has a set of N marbles in 4 colors: $m_0(1)$ green, $m_1(\mathrm{i})$ blue, $m_2(-1)$ yellow, and $m_3(-\mathrm{i})$ red ones; the marbles of the same color are assumed for the moment to be identical. Also, there are boxes of various sizes with $n+1$ compartments, which are consecutively labeled by 0, 1, 2, ..., n. The question is how to select a box of minimal size in order to arrange in it, one by one, all the N marbles, taking into account that the green marbles may be placed only into compartments, marked as 0, 4, 8, ..., blue marbles—into compartments, marked as 1, 5, 9, ..., yellow marbles—into compartments, marked as 2, 6, 10, ..., and red ones—into compartments 3, 7, 11, ...

To solve this rather simple combinatorial problem, observe that all odd dimensions $N \ge 3$ can be divided into two sets as:

$$a) \quad N = 3\,(mod\,4), \quad i.e., \quad N = 3 + 4l, \quad l = 0, 1, 2, 3, ...,$$
$$m_0(1) = m_1(i) = m_2(-1) = l + 1, \qquad m_3(-i) = l,$$
$$\sum_{k=0}^{3} m_k(\mathrm{i}^k) \;=\; 3(l+1) + l \;=\; 4l + 3 \;=\; N.$$

In particular, for $N = 3$ (i.e., $l = 0$) this set can be schematically depicted as

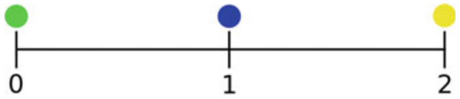


Fig. 1

$$b)\quad N = 5\,(mod\,4),\quad i.e.,\quad N = 5 + 4l,\quad l = 0, 1, 2, 3, ...,$$
$$m_0(1) = l + 2,\quad m_1(i) = m_2(-1) = m_3(-i) = l + 1,$$
$$\sum_{k=0}^{3} m_k(\mathrm{i}^k) \;=\; l + 2 +\; 3(l + 1) \;=\; 4l + 5 \;=\; N.$$

In particular, for $N = 5$ (i.e., $l = 0$) this set can be schematically depicted as

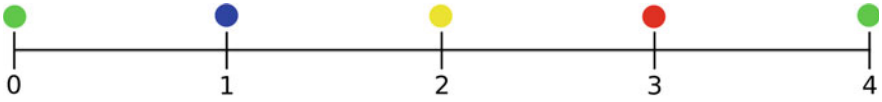


Fig. 2

By the same token, all even dimensions $N \geq 4$ can be divided into two sets of the form:

$$b)\quad N = 4\,(mod\,4),\quad i.e.,\quad N = 4 + 4l,\quad l = 0, 1, 2, 3, ...,$$
$$m_0(1) = l + 2,\quad m_1(i) = m_2(-1) = l + 1,\quad m_3(-i) = l,$$
$$\sum_{k=0}^{3} m_k(\mathrm{i}^k) \;=\; l + 2 +\; 2(l + 1) + l \;=\; 4(l + 1) \;=\; N\,;$$

In particular, for $N = 4$ (i.e., $l = 0$) this set can be schematically depicted as.

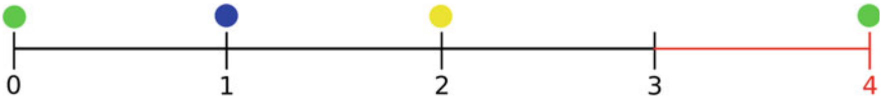


Fig. 3

$$b)\quad N = 6\,(mod\,4),\quad i.e.,\quad N = 6 + 4l,\quad l = 0, 1, 2, 3, ...,$$
$$m_0(1) = l + 2,\quad m_1(i) = l + 1,\quad m_2(-1) = l + 2,\quad m_3(-i) = l + 1,$$
$$\sum_{k=0}^{3} m_k(\mathrm{i}^k) \;=\; 2(l + 1) +\; 2(l + 2) \;=\; 4l + 6 \;=\; N.$$

In particular, for $N = 6$ (i.e., $l = 0$) this set can be schematically depicted as

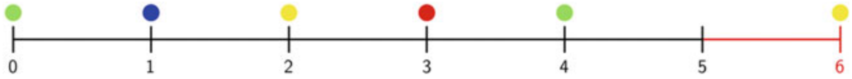


Fig. 4

Inspection of the 4 above figures indicates that for each odd number N of marbles it is sufficient to use a box, which contains only N compartments in it; whilst in the case of all even numbers N of marbles one needs to use boxes with $N+1$ compartments in them. Thus simply by examining the well-known formulas (4.2) for the multiplicities of the eigenvalues for the N-dimensional DFT operator $\Phi^{(N)}$ one arrives at the same conclusions concerning essential differences between symmetry properties of the eigenvectors for the even and odd dimensional DFT operator $\Phi^{(N)}$, as we elaborated at the end of the previous section.

5 Lowest Odd-Dimensional DFT Eigenvector

Now, before proceeding to the problem of finding the lowest DFT eigenvector $\mathbf{f}_0^{(N)}$ explicitly, we call attention to the following symmetry properties of the DFT eigenvectors $\mathbf{f}_n^{(N)}$, $0 \le n \le N-1$, in general. In the continuous case the number operator $\mathbf{N}$ commutes with the reflection operator $\mathbf{P}$, which is defined on the full real line $x \in \mathbb{R}$ as $\mathbf{P}\,x = -x$. Hence the Hermite functions (1.2) are either reflection symmetric or antisymmetric, that is, $\mathbf{P}\,\psi_n(x) = \psi_n(-x) = (-1)^n\,\psi_n(x)$. The discrete analogue of the reflection operator $\mathbf{P}$ is represented by the $N \times N$ matrix

$$\mathbf{P}_d := \mathbf{V}\,\mathbf{J} \equiv \mathbf{J}\,\mathbf{V}^{\dagger}, \tag{5.1}$$

where $\mathbf{J}$ is the $N \times N$ matrix with ones on the secondary diagonal, $\mathbf{V} = \mathbf{T}^{(-)}$ and $\mathbf{V}^{\dagger} = \mathbf{T}^{(+)}$ (recall that $T_{kl}^{(\pm)} = \delta_{k\pm1,l}$, $0 \le k, l \le N-1$). Similar to the continuous case, the discrete number operator $\mathscr{N}^{(N)}$ is $\mathbf{P}_d$ -symmetric, that is, $[\,\mathscr{N}^{(N)}, \mathbf{P}_d\,] = 0$.

Therefore it is plain that the DFT eigenvectors $\mathbf{f}_n^{(N)}$, $0 \le n \le N-1$, should be either $\mathbf{P}_d$ -symmetric or $\mathbf{P}_d$ -antisymmetric: from $\mathbf{P}_d\,\mathbf{f}_n^{(N)} = \mathbf{g}_n^{(N)}$ it follows that $\left(\mathbf{g}_n^{(N)}\right)_k = (-1)^n \left(\mathbf{f}_n^{(N)}\right)_{N-k-1}$.

It is natural to define the lowest DFT eigenvector $\mathbf{f}_0^{(N)}$ by the difference equation $\mathbf{b}_N\,\mathbf{f}_0^{(N)} = 0$. In other words, the lowest DFT eigenvector $\mathbf{f}_0^{(N)}$ is defined as the eigenvector of the lowering difference operator $\mathbf{b}_N$, associated with the zero eigenvalue. But as we have argued in Sect. 3, the null space of the matrix $\mathbf{b}_N$ is one-dimensional for all DFT's, based on odd points $N = 2L+1$, whereas it is two-dimensional for all DFT's on even points $N = 2L$. This means that the lowest DFT eigenvector $\mathbf{f}_0^{(N)}$ of the discrete number operator $\mathscr{N}^{(N)}$ is uniquely defined for all odd dimensions $N = 2L+1$, whereas in the cases of even dimensions $N = 2L$ the $\mathbf{P}_d$ -symmetry in the space of eigenvectors of the operator $\mathscr{N}^{(N)}$ is spontaneously broken. Hence only for odd dimensions $N = 2L+1$ it is possible to construct a ladder-type hierarchy explicitly for the all higher DFT eigenvectors $\mathbf{f}_n^{(N)}$, $1 \le n \le N-1 = 2L$, of the form

$$\mathbf{f}_n^{(N)} = \frac{1}{d_n}\left(\mathbf{b}_N^{\mathsf{T}}\right)^n \mathbf{f}_0^{(N)}, \qquad d_n := (\lambda_1\lambda_2...\lambda_n)^{1/2}, \tag{5.2}$$

where $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of the discrete number operator $\mathscr{N}^{(N)}$. So to begin with, one needs to find first the lowest DFT eigenvector $\mathbf{f}_0^{(N)}$ analytically.

Let us look for solutions of the difference equation $\mathbf{b}_N\, \mathbf{f}_0^{(N)} = 0$ in the form

$$\mathbf{f}_0^{(N)} = (x_0, x_1, x_2, \ldots, x_{2L-1}, x_{2L})^T, \qquad N = 2L+1, \tag{5.3}$$

without assuming *a priori* knowledge of whether this solution is symmetric or antisymmetric. Then the equation $\mathbf{b}_N\, \mathbf{f}_0^{(N)} = 0$ reduces to the following $N = 2L+1$ homogeneous equations

$$\begin{gathered} x_1 - x_{2L} = 0, \\ s_1x_1 + x_2 - x_0 = 0, \\ s_2x_2 + x_3 - x_1 = 0, \\ \cdots \quad \cdots \\ s_Lx_L + x_{L+1} - x_{L-1} = 0, \\ -s_Lx_{L+1} + x_{L+2} - x_L = 0, \\ \cdots \quad \cdots \\ -s_2x_{2L-1} + x_{2L} - x_{2L-2} = 0, \\ -s_1x_{2L} + x_0 - x_{2L-1} = 0, \end{gathered} \tag{5.4}$$

for the $N = 2L+1$ components $x_0, x_1, x_2, \ldots, x_{2L-1}, x_{2L}$ of the vector $\mathbf{f}_0^{(N)}$. From the first line in (5.4) it follows at once that $x_{2L} = x_1$, from the second and last lines it follows that $x_{2L-1} = x_2$, and so on. This means that the vector $\mathbf{f}_0^{(N)}$ has only $L+1$ linearly independent components and can be written as

$$\mathbf{f}_0^{(N)} = (x_0, x_1, x_2, \ldots, x_L, x_L, \ldots, x_2, x_1)^T, \tag{5.5}$$

whereas the system (5.4) reduces to the system of L homogeneous equations

$$\begin{gathered} x_0 - s_1x_1 - x_2 = 0, \\ x_1 - s_2x_2 - x_3 = 0, \\ x_2 - s_3x_3 - x_4 = 0, \\ \cdots \quad \cdots \\ x_{L-3} - s_{L-2}x_{L-2} - x_{L-1} = 0, \\ x_{L-2} - s_{L-1}x_{L-1} - x_L = 0, \\ x_{L-1} - s_Lx_L - x_L = 0, \end{gathered} \tag{5.6}$$

for those $L+1$ independent components $x_0, x_1, x_2, \ldots, x_L$ of the vector $\mathbf{f}_0^{(N)}$.

It is worth noting that from (5.5) it is already evident that the lowest odd-dimensional DFT eigenvector $\mathbf{f}_0^{(N)}$ is $\mathbf{P}_d$-symmetric, $\mathbf{P}_d\, \mathbf{f}_0^{(N)} = \mathbf{f}_0^{(N)}$ for all

$N = 2L + 1$, and this symmetry property of $\mathbf{f}_0^{(N)}$ follows directly from the defining difference equation $\mathbf{b}_N\, \mathbf{f}_0^{(N)} = 0$ itself.

To solve the system (5.6), one of the components $x_0, x_1, x_2, \ldots, x_L$ should be considered as an arbitrary one (note that from the last line in (5.6) it is clear that it should be the component x_L), so that the system (5.6) transforms into a system of L inhomogeneous equations

$$\begin{gathered}
x_0 - s_1 x_1 - x_2 = 0, \\
x_1 - s_2 x_2 - x_3 = 0, \\
x_2 - s_3 x_3 - x_4 = 0, \\
\cdots \qquad \cdots \\
x_{L-3} - s_{L-2} x_{L-2} - x_{L-1} = 0, \\
x_{L-2} - s_{L-1} x_{L-1} - x_L = 0, \\
x_{L-1} = (1 + s_L)\, x_L,
\end{gathered} \tag{5.7}$$

for the L components $x_0, x_1, x_2, \ldots, x_{L-1}$.

The augmented matrix associated with this linear system of equations is

$$\left(\begin{array}{cccccccc|c}
1 & -s_1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & -s_2 & -1 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 1 & -s_3 & -1 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & -s_{L-3} & -1 & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1 & -s_{L-2} & -1 & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 1 & -s_{L-1} & 1 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 & s_L + 1
\end{array}\right). \tag{5.8}$$

Since the matrix coefficient in (5.8) is represented by an upper triangular matrix with units on its diagonal, its determinant is equal to 1. Therefore, by employing Cramer's rule one obtains all unknowns x_k, $0 \le k \le L - 1$, in terms of the following $L \times L$ determinants:

$$x_k = \left|\begin{array}{ccccc:ccccc}
1 & -s_1 & -1 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & -s_2 & \ddots & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & 1 & \ddots & -1 & 0 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & -s_{k-1} & 0 & 0 & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 & -1 & 0 & \cdots & 0 \\
\hdashline
0 & 0 & 0 & \cdots & 0 & 0 & -s_{k+1} & -1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & -s_{L-2} & -1 \\
0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & 1 & -s_{L-1} \\
0 & 0 & 0 & \cdots & 0 & s_L + 1 & \cdots & 0 & 0 & 1
\end{array}\right|. \tag{5.9}$$

Note that the first principal $(k-1)\times(k-1)$ minor in (5.9), delineated by the dashed line in the left upper part of the determinant, is also of the upper triangular form and its determinant is therefore equal to 1. Hence, each unknown x_k can be also expressed in terms of the $(L-k)\times(L-k)$ determinants of the form

$$x_k = \begin{vmatrix} 0 & -s_{k+1} & -1 & \cdots 0 & 0 & 0 & 0 \\ 0 & 1 & -s_{k+2} & \cdots 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots 1 & -s_{L-3} & -1 & 0 \\ 0 & 0 & 0 & \cdots 0 & 1 & -s_{L-2} & -1 \\ 1 & 0 & 0 & \cdots 0 & 0 & 1 & -s_{L-1} \\ 1+s_L & 0 & 0 & \cdots 0 & 0 & 0 & 1 \end{vmatrix}. \tag{5.10}$$

This means that only the unknown x_0 is actually determined as the $L\times L$ determinant, defined by (5.10) with $k=0$, whereas all the other unknowns x_k, $1\le k\le L-1$, are expressed in terms of the descending $(L-k)\times(L-k)$ determinants (5.10), respectively.

Thus we conclude that the normalized lowest odd-dimensional eigenvector $\mathbf{f}_0^{(N)}$ of the discrete number operator $\mathscr{N}^{(N)}$ can be explicitly written as

$$\mathbf{f}_0^{(N)} = c_0^{-1}\,(x_0, x_1, \ldots, x_{L-1}, 1, 1, x_{L-1}, \ldots, x_1), \tag{5.11}$$

where all L independent components x_k, $k=0,\ldots,L-1$, are defined by (5.10) in terms of the $(L-k)\times(L-k)$ determinants, respectively, and the normalization constant

$$c_0 = \left(\frac{2\pi}{N}\right)^{1/4}\left[x_0^2 + 2\left(1 + x_1^2 + \ldots + x_{L-1}^2\right)\right]^{1/2}. \tag{5.12}$$

In connection with the formula (5.11), it may be remarked that from the algebraic point of view this procedure of deriving the explicit form of the lowest eigenvector $\mathbf{f}_0^{(N)}$ is equivalent to finding a basis for the null space of the matrix $\mathbf{b}_N$.

Recall that the null space of the matrix $\mathbf{b}_N$ is the subspace of $\mathbb{R}^N$ consisting of all solutions of the linear system $\mathbf{b}_N\,\mathbf{f}_0^{(N)} = 0$ (see, for example, [15]). To solve this system, one can put $\mathbf{b}_N$ in reduced echelon form by using row operations in the same way as we detailed it in Sect. 3. Then this system $\mathbf{b}_N\,\mathbf{f}_0^{(N)} = 0$ can be solved quickly by the process of back substitution, to get the same solution $\mathbf{f}_0^{(N)}$ as in (5.11), up to a constant factor.

Note that more convenient explicit expressions for the independent components x_k of the lowest eigenvector $\mathbf{f}_0^{(N)}$ may be found by using the Laplacian determinant expansion by the minors, associated with the only two nonzero elements in the first column in (5.10), in order to reduce this determinant to the sum of two $(L-k-2)\times(L-k-2)$ and $(L-k-1)\times(L-k-1)$ three-diagonal determinants of the same type,

$$x_k = \varDelta_{L-k-2}(s_{k+1}, \dots, s_{L-2}) + (1 + s_L)\varDelta_{L-k-1}(s_{k+1}, \dots, s_{L-1}). \tag{5.13}$$

Observe that in (5.13) we have used the notation

$$\varDelta_n(a_1, \dots, a_n) := \begin{vmatrix} a_1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & a_2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & a_3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & a_{n-2} & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & a_{n-1} & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & a_n \end{vmatrix}, \quad n \geq 1, \quad \varDelta_0 = 1, \tag{5.14}$$

for the determinant, which satisfies two three-term recursions

$$\begin{aligned} \varDelta_n(a_1, ..., a_n) &= a_1\, \varDelta_{n-1}(a_2, ..., a_n) + \varDelta_{n-2}(a_3, ..., a_n), \\ \varDelta_n(a_1, ..., a_n) &= a_n\, \varDelta_{n-1}(a_1, ..., a_{n-1}) + \varDelta_{n-2}(a_1, ..., a_{n-2}). \end{aligned} \tag{5.15}$$

The above recursions are readily derived from (5.14) with the aid of the Laplacian determinant expansion by the minors, associated with the only two nonzero elements either in the first column in (5.14), or in the last column of (5.14), respectively. Hence an explicit dependence of the determinant $\varDelta_n(a_1, ..., a_n)$ on the parameters $a_1, ..., a_n$ can be formulated in terms of the particular finite continued fractions, containing those parameters $a_1, ..., a_n$. Indeed, upon employing the square bracket notation for the finite continued fraction (see, for example, [16])

$$[\![a_1, a_2, \dots, a_n]\!] := a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}, \tag{5.16}$$

the determinant $\varDelta_n(a_1, ..., a_n)$ can be represented by a product of finite continued fractions as follows.

Lemma 2. *Let $a_1, a_2, \dots, a_n$ be real numbers. Then the determinant $\varDelta_n(a_1, ..., a_n)$ can be written in the form*

$$\varDelta_n(a_1, \dots, a_n) = [\![a_n, \dots, a_1]\!][\![a_{n-1}, \dots, a_1]\!] \cdots [\![a_2, a_1]\!][\![a_1]\!], \tag{5.17}$$

which is valid for $\forall n \in \mathbb{N}$.

Proof. We proceed by induction. For $n = 1$, $\varDelta_1(a_1) = a_1 = [\![a_1]\!]$. Now suppose that equality (5.17) holds for all $1 < n \leq m$. Then for $n = m + 1$, from the second line in (5.15) and the induction hypothesis one obtains that

$$\begin{aligned}\Delta_{m+1}(a_1,\ldots,a_{m+1}) &= a_{m+1}\Delta_m(a_1,\ldots,a_m)+\Delta_{m-1}(a_1,\ldots,a_{m-1})\\ = a_{m+1}[\![a_m,\ldots,a_1]\!][\![a_{m-1},\ldots,a_1]\!]&\cdots[\![a_2,a_1]\!][\![a_1]\!]+[\![a_{m-1},\ldots,a_1]\!]\cdots[\![a_2,a_1]\!][\![a_1]\!]\\ &= \Big(a_{m+1}[\![a_m,\ldots,a_1]\!]+1\Big)[\![a_{m-1},\ldots,a_1]\!]\cdots[\![a_2,a_1]\!][\![a_1]\!]\\ = [\![a_{m+1},a_m,\ldots,a_1]\!][\![a_m,\ldots,a_1]\!]&[\![a_{m-1},\ldots,a_1]\!]\cdots[\![a_2,a_1]\!][\![a_1]\!],\end{aligned} \tag{5.18}$$

where in the penultimate line the identity

$$a_{m+1}[\![a_m,\ldots,a_1]\!]+1=[\![a_{m+1},a_m,\ldots,a_1]\!]\,[\![a_m,\ldots,a_1]\!] \tag{5.19}$$

is employed, which represents just another form of writing down the definition (5.16) for the bracket notation. □

We are now in a position to express the components of the lowest odd-dimensional eigenvector $\mathbf{f}_0^{(N)}$ in terms of finite continued fractions as follows.

Proposition 2. *Let* $\mathbf{b}_N$ *be the matrix given by (2.20) and* $\mathbf{f}_0^{(N)}$ *be the solution of the difference equation* $\mathbf{b}_N\,\mathbf{f}_0^{(N)}=0$*, given by (5.11). Then* $x_{L-1}=1+s_L$ *and all other components* x_k*,* $k=0,\ldots,L-2$*, are equal to*

$$x_k=\Big(1+[\![s_L,s_{L-1},\ldots,s_{k+2},s_{k+1}]\!]\Big)\prod_{j=0}^{L-k-2}[\![s_{L-1-j},s_{L-2-j},\ldots,s_{k+2},s_{k+1}]\!]. \tag{5.20}$$

Proof. It follows from (5.7) that $x_{L-1}=1+s_L$. By using the identity

$$s_L\,[\![s_{L-1},s_{L-2},\ldots,s_{k+1}]\!]=[\![s_L,s_{L-1},\ldots,s_{k+1}]\!]\,[\![s_{L-1},s_{L-2},\ldots,s_{k+1}]\!]-1, \tag{5.21}$$

which is valid by (5.18), from (5.13) and the above lemma one derives that

$$\begin{aligned}x_k &= [\![s_{L-2},\ldots,s_{k+1}]\!]\,[\![s_{L-3},\ldots,s_{k+1}]\!]\cdots[\![s_{k+2},s_{k+1}]\!]\,[\![s_{k+1}]\!]\\ &+(s_L+1)[\![s_{L-1},\ldots,s_{k+1}]\!]\,[\![s_{L-2},\ldots,s_{k+1}]\!]\cdots[\![s_{k+2},s_{k+1}]\!]\,[\![s_{k+1}]\!]\\ &=[\![s_L,s_{L-1},\ldots,s_{k+2},s_{k+1}]\!]\cdots[\![s_{k+1}]\!]+[\![s_{L-1},s_{L-2},\ldots,s_{k+2},s_{k+1}]\!]\cdots[\![s_{k+1}]\!]\\ &=\Big(1+[\![s_L,s_{L-1},\ldots,s_{k+2},s_{k+1}]\!]\Big)\prod_{j=0}^{L-k-2}[\![s_{L-1-j},s_{L-2-j},\ldots,s_{k+2},s_{k+1}]\!],\end{aligned} \tag{5.22}$$

where $k=0,\ldots,L-2$. This completes the proof of (5.20). □

6 Concluding Remarks

To summarize, we have thus succeeded in deriving the discrete number operator $\mathcal{N}^{(N)}$ in terms of the raising and lowering difference operators, which satisfy the standard intertwining with the DFT operator. We proved that the discrete number operator $\mathscr{N}^{(N)}$ has distinct eigenvalues only if this operator is associated with the DFT's based on grids $\{x_0, x_1, ..., x_{N-1}\}$ with odd $N = 2L + 1$, whereas in the cases of the DFT's based on grids $\{x_0, x_1, ..., x_{N-1}\}$ with even $N = 2L$ the discrete reflection symmetry in the space of eigenvectors of the discrete number operator $\mathscr{N}^{(N)}$ is spontaneously broken. This essential distinction between even and odd dimensions has been shown to be consistent with the well-known old formula for the multiplicities of the eigenvalues, associated with the N-dimensional DFT. Finally, we have explicitly found the lowest DFT eigenvector $\mathbf{f}_0^{(N)}$ of the operator $\mathscr{N}^{(N)}$ for all odd dimensions $N = 2L + 1$, which helped us to formulate a ladder-type hierarchy of all the higher DFT eigenvectors $\mathbf{f}_n^{(N)}$, $1 \leq n \leq N - 1 = 2L$. However, what is missing in this exposition of our current understanding of the discrete Fourier transform, is the explicit form of the discrete analogue of the eigenfunctions of the Fourier integral transform in terms of the Hermite polynomials $H_n(x)$ times the lowest FIT eigenfunction $\psi_0(x) = e^{-x^2/2}$. We believe that we need just a bit more time in order to resolve this final piece of the puzzle, which has for a long time surrounded the explicit form of the eigenvectors of the discrete Fourier transform.

Acknowledgements We are grateful to Fernanado González for graphical support of our results. M.K.A. thanks the support of Proyecto CONACYT No. A1-S-8793 "Problemas de frontera para ecuaciones integrales y diferenciales parciales de análisis y física matemática" (Universidad Autónoma del Estado de Morelos), N.M.A. thanks Project No. IG-100119 awarded by the Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México.

References

1. Landau, L.D., Lifshitz, E.M.: Quantum Mechanics (Non-relativistic Theory). Pergamon Press, Oxford (1991)
2. McClellan, J.H., Parks, T.W.: Eigenvalue and eigenvector decomposition of the discrete Fourier transform. IEEE Trans. Audio Electroacoust. **AU-20**, 66–74 (1972)
3. Auslander, L., Tolimieri, R.: Is computing with the finite Fourier transform pure or applied mathematics? Bull. Am. Math. Soc. **1**, 847–897 (1979)
4. Dickinson, B.W., Steiglitz, K.: Eigenvectors and functions of the discrete Fourier transform. IEEE Trans. Acoust. Speech Signal Process. **30**, 25–31 (1982)
5. Mehta, M.L.: Eigenvalues and eigenvectors of the finite Fourier transform. J. Math. Phys. **28**, 781–785 (1987)
6. Atakishiyev, N.M.: On q-extensions of Mehta's eigenvectors of the finite Fourier transform. Int. J. Mod. Phys. A **21**, 4993–5006 (2006)
7. Atakishiyeva, M.K., Atakishiyev, N.M.: On the raising and lowering difference operators for eigenvectors of the finite Fourier transform. J. Phys. Conf. Ser. **597**, 012012 (2015)
8. Atakishiyeva, M.K., Atakishiyev, N.M.: On algebraic properties of the discrete raising and lowering operators, associated with the N-dimensional discrete Fourier transform. Adv. Dyn. Syst. Appl. **11**, 81–92 (2016)
9. Atakishiyeva, M.K., Atakishiyev, N.M., Méndez Franco, J.: On a discrete number operator associated with the 5D discrete Fourier transform. In: Differential and Difference Equations with Applications. Springer Proceedings in Mathematics & Statistics, vol. 164, pp. 273–292 (2016)
10. Atakishiyeva, M.K., Atakishiyev, N.M., Loreto-Hernández, J.: More on algebraic properties of the discrete Fourier transform raising and lowering operators. 4 Open, vol. 2, pp. 1–11 (2019)

11. Mugler, D.H., Clary, S.: Discrete Hermite functions. In: Proceedings of the International Conference on Scientific Computing and Mathematical Modeling, Milwaukee WI, pp. 318–321 (2000)
12. Mugler, D.H., Clary, S.: Discrete Hermite functions and the fractional Fourier transform. In: Proceedings of the International Conference on Sampling Theory and Applications, Orlando FL, pp. 303–308 (2001)
13. Clary, S., Mugler, D.H.: Shifted Fourier matrices and their tridiagonal commutators. SIAM J. Matrix Anal. Appl. **24**, 809–821 (2003)
14. Shapiro, H.: Linear Algebra and Matrices. AMS, Providence, Rhode Island (2015)
15. Robinson, D.J.S.: A Course in Linear Algebra with Applications. World Scientific, Singapore (2006)
16. Hardy, G.H., Wright, E.M.: An Introduction to the Theory of Numbers. Oxford University Press, Oxford (2008)

General Sets of Bell-Sheffer and Log-Sheffer Polynomials

Pierpaolo Natalini, Sandra Pinelas, and Paolo Emilio Ricci

Abstract The introduction of iterated exponential and logarithmic functions allows to construct new sets of Bell-Sheffer and logarithmic Sheffer (shortly log-Sheffer) polynomials, whose shift operators and differential equations exhibit an iterative character. In this context it is possible to define, for every integers r and s, polynomials of higher order. They give back, in particular cases, the Bell-exponential and logarithmic polynomials and numbers introduced in preceding papers. Connections with integer sequences appearing in Combinatorial analysis are also mentioned.

Keywords Sheffer polynomials · Generating functions · Monomiality principle · Shift operators · Combinatorial analysis

AMS 2010 Mathematics Subject Classifications. 33C99 · 05A10 · 11P81

1 Introduction

In recent articles [8, 9, 21], new sets of Sheffer [26] and Brenke [7] polynomials, based on higher order Bell numbers [4, 13, 15, 16, 21], have been studied. Furthermore, several integer sequences associated [27] with the considered polynomials sets both of exponential [1, 2] and logarithmic [8] type have been introduced.

It is worth to note that exponential and logarithmic polynomials have been recently studied in the multidimensional case [18–20].

P. Natalini (✉)
Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre,
Largo San Leonardo Murialdo, 1, 00146 Roma, Italy
e-mail: natalini@mat.uniroma3.it

S. Pinelas
Departamento de Ciências Exactas e Engenharia, Academia Militar,
Av. Conde Castro Guimarães, 2720-113 Amadora, Portugal
e-mail: sandra.pinelas@gmail.com

P. E. Ricci
International Telematic University UniNettuno, Corso Vittorio Emanuele II, 39, 00186 Roma, Italy
e-mail: paoloemilioricci@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_42

In preceding articles [16, 22] some particular cases of exponential-Sheffer and logarithmic-Sheffer polynomials have been introduced. In this article we extend these results to the general case of exponential (Bell-Sheffer) and logarithmic-Sheffer (Log-Sheffer) polynomials.

2 Sheffer Polynomials

The Sheffer polynomials $\{s_n(x)\}$ are introduced [26] by means of the exponential generating function [28] of the type:

$$A(t)\exp(xH(t)) = \sum_{n=0}^{\infty} s_n(x)\,\frac{t^n}{n!}, \tag{2.1}$$

where

$$\begin{aligned} A(t) &= \sum_{n=0}^{\infty} a_n\,\frac{t^n}{n!}\,, \qquad (a_0 \neq 0),\\ H(t) &= \sum_{n=0}^{\infty} h_n\,\frac{t^n}{n!}\,, \qquad (h_0 = 0). \end{aligned} \tag{2.2}$$

According to a different characterization (see [25, p. 18]), the same polynomial sequence can be defined by means of the pair $(g(t), f(t))$, where $g(t)$ is an invertible series and $f(t)$ is a delta series:

$$\begin{aligned} g(t) &= \sum_{n=0}^{\infty} g_n\,\frac{t^n}{n!}\,, \qquad (g_0 \neq 0),\\ f(t) &= \sum_{n=0}^{\infty} f_n\,\frac{t^n}{n!}\,, \qquad (f_0 = 0,\, f_1 \neq 0). \end{aligned} \tag{2.3}$$

Denoting by $f^{-1}(t)$ the compositional inverse of $f(t)$ (i.e. such that $f\left(f^{-1}(t)\right) = f^{-1}\left(f(t)\right) = t$), the exponential generating function of the sequence $\{s_n(x)\}$ is given by

$$\frac{1}{g[f^{-1}(t)]}\exp\left(xf^{-1}(t)\right) = \sum_{n=0}^{\infty} s_n(x)\,\frac{t^n}{n!}, \tag{2.4}$$

so that

$$A(t) = \frac{1}{g[f^{-1}(t)]}, \qquad H(t) = f^{-1}(t). \tag{2.5}$$

When $g(t) \equiv 1$, the Sheffer sequence corresponding to the pair $(1, f(t))$ is called the associated Sheffer sequence $\{\sigma_n(x)\}$ for $f(t)$, and its exponential generating function is given by

$$\exp\left(xf^{-1}(t)\right) = \sum_{n=0}^{\infty} \sigma_n(x) \frac{t^n}{n!}. \tag{2.6}$$

A list of known Sheffer polynomial sequences and their associated ones can be found in [5, 6].

2.1 Shift Operators and Differential Equation

We recall that a polynomial set $\{p_n(x)\}$ is called quasi-monomial if and only if there exist two operators $\hat{P}$ and $\hat{M}$ such that

$$\hat{P}\,(p_n(x)) = np_{n-1}(x), \qquad \hat{M}\,(p_n(x)) = p_{n+1}(x), \qquad (n = 1, 2, \ldots). \tag{2.7}$$

$\hat{P}$ is called the *derivative* operator and $\hat{M}$ the *multiplication* operator, as they act in the same way of classical operators on monomials.

This definition traces back to a paper by Steffensen [29] recently improved by Dattoli [10, 11] and widely used in several applications [12].

Ben Cheikh [3] proved that every polynomial set is quasi-monomial under the action of suitable derivative and multiplication operators. In particular, in the same article, the following result is proved, as a particular case of Corollary 3.2:

Theorem 2.1. *Let $(p_n(x))$ denote a Sheffer polynomial set, defined by the generating function*

$$A(t)\exp(xH(t)) = \sum_{n=0}^{\infty} p_n(x) \frac{t^n}{n!}, \tag{2.8}$$

where

$$A(t) = \sum_{n=0}^{\infty} \tilde{a}_n t^n, \qquad (\tilde{a}_0 \neq 0), \tag{2.9}$$

and

$$H(t)=\sum_{n=0}^{\infty}\tilde{h}_n\,t^{n+1},\qquad (\tilde{h}_0\neq 0). \tag{2.10}$$

Denoting, as before, by $f(t)$ the compositional inverse of $H(t)$, the Sheffer polynomial set $\{p_n(x)\}$ is quasi-monomial under the action of the operators

$$\hat{P}=f(D_x),\qquad \hat{M}=\frac{A'[f(D_x)]}{A[f(D_x)]}+xH'[f(D_x)], \tag{2.11}$$

where prime *denotes the ordinary derivatives with respect to t.*

Furthermore, according to the monomiality principle, the quasi-monomial polynomials $\{p_n(x)\}$ satisfy the differential equation

$$\hat{M}\hat{P}\,p_n(x)=n\,p_n(x). \tag{2.12}$$

3 New Exponential and Logarithmic-Sheffer Polynomial Sets

We introduce, for shortness, the following compact notation.

Put, by definition:

$E_0(t) := \exp(t)-1$
$E_1(t) := E_0(E_0(t)) = \exp(\exp(t)-1)-1$
....
$E_r(t) := E_0(E_{r-1}(t)) = \exp(\exp(\dots(\exp(t)-1)\dots)-1)-1,\quad [(r+1)-\text{times}\ \ \exp],$

$E_r(E_s(t)) = E_{r+s+1}(t),$

and in a similar way:

$\Lambda_0(t) := \log(t+1)$
$\Lambda_1(t) := \Lambda_0(\Lambda_0(t)) = \log(\log(t+1)+1)$
....
$\Lambda_r(t) := \Lambda_0(\Lambda_{r-1}(t)) = \log\left(\log\left(\dots(\log(t+1)+1)\dots\right)+1\right),\quad [(r+1)-\text{times}\ \ \log],$

$\Lambda_r(\Lambda_s(t)) = \Lambda_{r+s+1}(t).$

3.1 Operational Rules

Note that, for every integers r, k, h,

$$E_r(\Lambda_r(t)) = t, \qquad \Lambda_r(E_r(t)) = t,$$

$$\begin{aligned} &(\mathit{if}\ k > h) \quad E_k(\Lambda_h(t)) = E_{k-h-1}(t), \quad E_h(\Lambda_k(t)) = \Lambda_{k-h-1}(t), \\ &(\mathit{if}\ k > h) \quad \Lambda_k(E_h(t)) = \Lambda_{k-h-1}(t), \quad \Lambda_h(E_k(t)) = E_{k-h-1}(t), \end{aligned} \tag{3.1}$$

$$e^{E_r(t)} = E_{r+1}(t) + 1, \qquad e^{\Lambda_r(t)} = \Lambda_{r-1}(t) + 1,$$

and it is suitable to put, by definition:

$$E_{-1}(t) := \Lambda_{-1}(t) := t. \tag{3.2}$$

Furthermore, the differentiation rules hold

- For the exponential functions

$$\begin{aligned} D_t E_0(t) &= e^t = E_0(t) + 1, \\ D_t E_1(t) &= [E_1(t) + 1][E_0(t) + 1], \end{aligned} \tag{3.3}$$

  and, in general, for every $r \geq 0$,

$$D_t E_r(t) = \prod_{\ell=0}^{r} [E_\ell(t) + 1], \tag{3.4}$$

- For the logarithmic functions

$$\begin{aligned} D_t \Lambda_0(t) &= \frac{1}{t+1}, \\ D_t \Lambda_1(t) &= \frac{1}{[\log(t+1) + 1](t+1)} = \frac{1}{[\Lambda_0(t) + 1](t+1)}, \end{aligned} \tag{3.5}$$

  and in general, for every $s \geq 0$,:

$$D_t \Lambda_s(t) = \left[\prod_{\ell=0}^{s} \left[\Lambda_{\ell-1}(t) + 1 \right] \right]^{-1}. \tag{3.6}$$

Remark 3.1. Note that the coefficients of the Taylor expansion of $E_1(t)$ are given by the Bell numbers $b_n = b_n^{[1]}$

$$E_1(t) + 1 = \sum_{n=0}^{\infty} b_n^{[1]} \frac{t^n}{n!},$$

and, in general the coefficients of the Taylor expansion of $E_r(t)$ are given by the higher order Bell numbers $b_n^{[r]}$

$$E_r(t) + 1 = \sum_{n=0}^{\infty} b_n^{[r]} \frac{t^n}{n!}. \tag{3.7}$$

The higher order Bell numbers, also known as higher order exponential numbers, have been considered in [13–15], and used in [21] in the framework of Brenke and Sheffer polynomials. The first few of them are shown in Table 1.

Remark 3.2. Note that the coefficients of the Taylor expansion of $\Lambda_0(t)$ are given by the logarithmic numbers $l_n^{[1]} = (-1)^{n-1}(n-1)!$

$$\Lambda_0(t) = \sum_{n=1}^{\infty} l_n^{[1]} \frac{t^n}{n!} = \sum_{n=1}^{\infty} (-1)^{n-1}(n-1)! \frac{t^n}{n!}, \tag{3.8}$$

and, in general the coefficients of the Taylor expansion of $\Lambda_{r-1}(t)$ are given by the higher order logarithmic numbers $l_n^{[r]}$

Table 1 Bell and higher order Bell numbers for $n = 1, 2, \ldots, 10$

n	$b_n^{[1]}$	$b_n^{[2]}$	$b_n^{[3]}$	$b_n^{[4]}$	$b_n^{[5]}$
0	1	1	1	1	1
1	1	1	1	1	1
2	2	3	4	5	6
3	5	12	22	35	51
4	15	60	154	315	561
5	52	358	1304	3455	7556
6	203	2471	12915	44590	120196
7	877	19302	146115	660665	2201856
8	4140	167894	1855570	11035095	45592666
9	21147	1606137	26097835	204904830	1051951026
10	115975	16733779	402215465	4183174520	26740775306

Table 2 Logarithmic numbers for $n = 1, 2, \dots, 10$

n	$l_n^{[1]}$	$l_n^{[2]}$	$l_n^{[3]}$	$l_n^{[4]}$	$l_n^{[5]}$
1	1	1	1	1	1
2	−1	−2	−3	−4	−5
3	2	7	15	26	40
4	−6	−35	−105	−234	−440
5	24	228	947	2696	6170
6	−120	−1834	−10472	−37919	−105315
7	720	17582	137337	630521	2120610
8	−5040	−195866	−2085605	−12111114	−49242470
9	40320	2487832	36017472	264051201	1296133195
10	−362880	−35499576	−697407850	−6445170229	−38152216495

$$\Lambda_{r-1}(t) = \sum_{n=1}^{\infty} l_n^{[r]} \frac{t^n}{n!}. \tag{3.9}$$

The higher order logarithmic numbers, which are the counterpart of the higher order Bell (exponential) numbers, have been considered in [8], and used there in the framework of new sets of Sheffer polynomials. The first few of them are shown in Table 2.

It is worth to note that the (absolute value) of numbers contained in Table 2, read by column for $r = 1, 2, \dots, 10$, are contained in the Encyclopedia of integer sequences [27], respectively under A000142, A003713, A000268, A000310, A000359.

The same table, read by rows, for $n = 3$ gives a symmetric integer sequence known under A005449 (second pentagonal numbers $n(3n + 1)/2$), for $n = 4$ gives a sequence known under A094952 (derived from pentagonal numbers, or from Stirling numbers of the first kind matrix), while the subsequent sequences for $n = 5, 6, \dots,$ are not included in the Encyclopedia.

Therefore, the above definitions are useful to clarify the meaning of the considered sequences.

Remark 3.3. In what follows, it is convenient to put, by definition, in (3.8)–(3.9), $l_0^{[r]} = 1$, $\forall r \geq 0$, so that $\Lambda_0(D_x) + 1 = \sum_{k=0}^{\infty} l_k^{[1]} t^n/n!$, and in general $\Lambda_{r-1}(D_x) + 1 = \sum_{k=0}^{\infty} l_k^{[r]} t^n/n!$.

3.2 *Cases to Be Considered*

In the following Sections we consider the Sheffer polynomial sets defined by the generating functions:

The Case Exp-Exp

- Exp-Exp I.1.

$$G(t,x) = \exp[E_r(t) + x\,E_s(t)] = \sum_{n=0}^{\infty} e_n^{[r,s]}(x)\,\frac{t^n}{n!}\,, \quad (r \geq s)$$

- Exp-Exp I.2.

$$G(t,x) = \exp[E_r(t) + x\,E_s(t)] = \sum_{n=0}^{\infty} \tilde{e}_n^{[r,s]}(x)\,\frac{t^n}{n!}\,, \quad (r < s)$$

The Case Log-Log

- Log-Log II.1.

$$G(t,x) = \exp[\Lambda_r(t) + x\,\Lambda_s(t)] = \sum_{n=0}^{\infty} \ell_n^{[r,s]}(x)\,\frac{t^n}{n!}, \quad (r \geq s)$$

- Log-Log II.2.

$$G(t,x) = \exp[\Lambda_r(t) + x\,\Lambda_s(t)] = \sum_{n=0}^{\infty} \tilde{\ell}_n^{[r,s]}(x)\,\frac{t^n}{n!}, \quad (r < s)$$

The Case Exp-Log

$$G(t,x) = \exp[E_r(t) + x\,\Lambda_s(t)] = \sum_{n=0}^{\infty} \varepsilon_n^{[r,s]}(x)\,\frac{t^n}{n!}, \quad (r \geq s)$$

The Case Log-Exp

$$G(t,x) = \exp[\Lambda_r(t) + x\,E_s(t)] = \sum_{n=0}^{\infty} \lambda_n^{[r,s]}(x)\,\frac{t^n}{n!}, \quad (r \geq s)$$

The particular cases Exp-Exp I.1 and Log-Log II.1, when $r = s$ have been already considered (see respectively [16, 22]). Other basic exponential and logarithmic Sheffer polynomial sets, defined by assuming $A(t) = e^t$ have been examined in preceding articles (see [17, 23]).

4 The Case Exp-Exp

We have:

$$A(t) = \exp[E_r(t)], \qquad H(t) = E_s(t),$$

$$\frac{A'(t)}{A(t)} = \prod_{\ell=0}^{r}[E_\ell(t) + 1], \qquad H'(t) = \prod_{\ell=0}^{s}[E_\ell(t) + 1],$$

$$H^{-1}(t) = \Lambda_s(t) = f(t),$$

so that

$$\hat{P} = \Lambda_s(D_x),$$

$$\hat{M} = \prod_{\ell=0}^{r}[E_\ell(\Lambda_s(D_x)) + 1] + x\prod_{\ell=0}^{s}[E_\ell(\Lambda_s(D_x)) + 1].$$

Remark 4.1. Note that, putting $a(t) = \exp[E_r(t)]$, $b(t) = \exp[E_s(t)]$ and $x = \theta$, the generating function writes:

$$G(t, x) = \exp[E_r(t)](\exp[E_s(t)])^x = a(t)[b(t)]^\theta$$

so that, for every fixed t, as a function of θ, is a logarithmic spiral [24].

4.1 *The Case Exp-Exp I.1*

Being $r \le s$, recalling the operational rules (3.1), we find:

$$\hat{M} = (1 + x)\prod_{\ell=0}^{r}[E_\ell(\Lambda_s(D_x)) + 1] + x\prod_{\ell=r+1}^{s}[E_\ell(\Lambda_s(D_x)) + 1] =$$

$$= (1 + x)\prod_{\ell=0}^{r}[\Lambda_{s-\ell-1}(D_x) + 1] + x\prod_{\ell=r+1}^{s}[\Lambda_{s-\ell-1}(D_x) + 1],$$

so that we have the differential equation:

$$\left\{(1 + x)\prod_{\ell=0}^{r}[\Lambda_{s-\ell-1}(D_x) + 1] + x\prod_{\ell=r+1}^{s}[\Lambda_{s-\ell-1}(D_x) + 1]\right\}\Lambda_s(D_x)\, e_n^{[r,s]}(x) = n\, e_n^{[r,s]}(x).$$

4.2 Example

Let $G(t,x) = \exp[E_1(t) + x\,E_2(t)] = \sum_{n=1}^{\infty} e_n^{[1,2]}(x)\,\frac{t^n}{n!}$.

The first few $e_n^{[1,2]}(x)$ polynomials are:

$$e_0^{[1,2]}(x) = 1$$

$$\begin{aligned}
e_1^{[1,2]}(x) &= x + 1 \\
e_2^{[1,2]}(x) &= x^2 + 5x + 3 \\
e_3^{[1,2]}(x) &= x^3 + 12x^2 + 30x + 12 \\
e_4^{[1,2]}(x) &= x^4 + 22x^3 + 129x^2 + 210x + 60 \\
e_5^{[1,2]}(x) &= x^5 + 35x^4 + 375x^3 + 1425x^2 + 1678x + 358 \\
e_6^{[1,2]}(x) &= x^6 + 51x^5 + 870x^4 + 6045x^3 + 16683x^2 + 15047x + 2471 \\
e_7^{[1,2]}(x) &= x^7 + 70x^6 + 1743x^5 + 19320x^4 + 97818x^3 + 208565x^2 + 149404x + 19302
\end{aligned}$$

Further values can be easily achieved by using Wolfram Alpha©.

Remark 4.2. Note that the sequence $\{1, 1, 3, 12, 60, 358, 2471, 19302, \dots\}$ appears in the Encyclopedia of Integer Sequences [27] under A000258—Number of 3-level labeled rooted trees with n leaves.—Christian G. Bower, Aug 15, 1998.

Differential Equation

$$\Big\{(1+x)\,[\Lambda_0(D_x)+1]\,[\Lambda_1(D_x)+1] + x\,(D_x+1)\Big\}\;\Lambda_2(D_x)\;e_n^{[1,2]}(x) = n\,e_n^{[1,2]}(x).$$

Therefore, using positions in Remark 3.3, we find:

$$\left[(1+x)\sum_{k=0}^{\infty} l_k^{[1]}\frac{D_x^k}{k!}\sum_{k=0}^{\infty} l_k^{[2]}\frac{D_x^k}{k!} + x\,(D_x+1)\right]\sum_{k=0}^{\infty} l_{k+1}^{[3]}\,\frac{D_x^{k+1}}{(k+1)!}\;\tilde{e}_n^{[2,1]}(x) = n\,\tilde{e}_n^{[2,1]}(x),$$

$$\Bigg[(1+x)\sum_{\substack{k_1+k_2+k_3=k\\ 0\le k\le n-1}}\binom{k}{k_1,k_2,k_3}\,l_{k_1}^{[1]}\,l_{k_2}^{[2]}\,l_{k_3+1}^{[3]}(k_3+1)\,D_x^{k+1}$$

$$+\;x\,(D_x+1)\sum_{k=0}^{n-1} l_{k+1}^{[3]}\,\frac{D_x^{k+1}}{(k+1)!}\Bigg]\;\tilde{e}_n^{[2,1]}(x) = n\,\tilde{e}_n^{[2,1]}(x),$$

since the series applied to a polynomial of degree n reduce to a finite sums.

4.3 The Case Exp-Exp I.2

Being $r > s$, recalling the operational rules (3.1), we find:

$$\begin{aligned}\hat{M} &= (1+x)\prod_{\ell=0}^{s}[E_\ell(\Lambda_s(D_x))+1] + \prod_{\ell=s+1}^{r}[E_\ell(\Lambda_s(D_x))+1] = \\ &= (1+x)\prod_{\ell=0}^{s}[\Lambda_{s-\ell-1}(D_x)+1] + \prod_{\ell=s+1}^{r}[E_{\ell-s-1}(D_x)+1],\end{aligned}$$

so that we have the differential equation:

$$\left\{(1+x)\prod_{\ell=0}^{s}[\Lambda_{s-\ell-1}(D_x)+1] + \prod_{\ell=s+1}^{r}[E_{\ell-s-1}(D_x)+1]\right\}\Lambda_s(D_x)\,\tilde{e}_n^{[r,s]}(x) = n\,\tilde{e}_n^{[r,s]}(x).$$

4.4 Example

Let $G(t,x) = \exp[E_2(t) + x\,E_1(t)] = \sum_{n=1}^{\infty}\tilde{e}_n^{[2,1]}(x)\,\dfrac{t^n}{n!}$.

The first few $\tilde{e}_n^{[2,1]}(x)$ polynomials are:

$$\tilde{e}_0^{[2,1]}(x) = 1$$

$$\tilde{e}_1^{[2,1]}(x) = x+1$$

$$\tilde{e}_2^{[2,1]}(x) = x^2+4x+4$$

$$\tilde{e}_3^{[2,1]}(x) = x^3+9x^2+23x+22$$

$$\tilde{e}_4^{[2,1]}(x) = x^4+16x^3+80x^2+171x+154$$

$$\tilde{e}_5^{[2,1]}(x) = x^5+25x^4+210x^3+795x^2+1537x+1304$$

$$\tilde{e}_6^{[2,1]}(x) = x^6+36x^5+460x^4+2765x^3+8932x^2+16059x+12915$$

$$\tilde{e}_7^{[2,1]}(x) = x^7+49x^6+889x^5+7875x^4+38577x^3+112378x^2+190339x+146115$$

Further values can be easily achieved by using Wolfram Alpha©.

Remark 4.3. Note that the sequence $\{1, 1, 4, 22, 154, 1304, 12915, 146115, \dots\}$ appears in the Encyclopedia of Integer Sequences [27] under A000307—Number of 4-level labeled rooted trees with n leaves.

Differential Equation

$$\Big[(1+x)(D_x+1)[\Lambda_0(D_x)+1]+e^{D_x}\Big]\Lambda_1(D_x)\,\tilde{e}_n^{[2,1]}(x)=n\,\tilde{e}_n^{[2,1]}(x),$$

Therefore, using positions in Remark 3.3, we find:

$$\Big[(1+x)(D_x+1)\sum_{k=0}^{\infty}l_k^{[1]}\frac{D_x^k}{k!}\sum_{k=0}^{\infty}l_{k+1}^{[2]}\frac{D_x^{k+1}}{(k+1)!} + \sum_{k=0}^{\infty}\frac{D_x^k}{k!}\sum_{k=0}^{\infty}l_{k+1}^{[2]}\frac{D_x^{k+1}}{(k+1)!}\Big]\,\tilde{e}_n^{[2,1]}(x)=n\,\tilde{e}_n^{[2,1]}(x),$$

$$\Big[(1+x)(D_x+1)\sum_{k=0}^{n-1}\sum_{h=0}^{k}\frac{l_{k-h}^{[1]}\,l_{h+1}^{[2]}}{(k-h)!\,(h+1)!}D_x^{k+1} + \sum_{k=0}^{n-1}\sum_{h=0}^{k}\frac{l_{h+1}^{[2]}}{(k-h)!\,(h+1)!}D_x^{k+1}\Big]\,\tilde{e}_n^{[2,1]}(x)=n\,\tilde{e}_n^{[2,1]}(x),$$

since the series applied to a polynomial of degree n reduce to a finite sums.

5 The Case Log-Log

We have:

$$A(t)=\exp[\Lambda_r(t)],\qquad H(t)=\Lambda_s(t),$$

$$\frac{A'(t)}{A(t)}=\left[\prod_{\ell=0}^{r}[\Lambda_{\ell-1}(t)+1]\right]^{-1},\qquad H'(t)=\left[\prod_{\ell=0}^{s}[\Lambda_{\ell-1}(t)+1]\right]^{-1},$$

$$H^{-1}(t)=E_s(t)=f(t),$$

so that

$$\hat{P} = E_s(D_x),$$

$$\hat{M} = \left[\prod_{\ell=0}^{r}[\Lambda_{\ell-1}(E_s(D_x)) + 1]\right]^{-1} + x\left[\prod_{\ell=0}^{s}[\Lambda_{\ell-1}(E_s(D_x)) + 1]\right]^{-1}.$$

Remark 5.1. Note that, here and in what follows, we could write

$$A(t) = \exp[\Lambda_r(t)] = \Lambda_{r-1}(t) + 1 \quad \text{and} \quad G(t,x) = (\Lambda_{r-1}(t) + 1)(\Lambda_{s-1}(t) + 1)^x,$$

but we used the notation above to highlight, here and in what follows, the symmetry with respect to the Exp case.

The above notation for $G(t,x)$, putting $a(t) = \Lambda_{r-1}(t) + 1$, $b(t) = \Lambda_{s-1}(t) + 1$ and $x = \theta$, shows that the generating function, for every fixed t, as a function of θ, is a logarithmic spiral [24].

5.1 The Case Log-Log I.1

Being $r \leq s$, recalling the operational rules (3.1), we find:

$$\begin{aligned}\hat{M} &= (1+x)\left[\prod_{\ell=0}^{r}[\Lambda_{\ell-1}(E_s(D_x)) + 1]\right]^{-1} + x\left[\prod_{\ell=r+1}^{s}[\Lambda_{\ell-1}(E_s(D_x)) + 1]\right]^{-1}\\ &= (1+x)\left[\prod_{\ell=0}^{r}[E_{s-\ell}(D_x) + 1]\right]^{-1} + x\left[\prod_{\ell=r+1}^{s}[E_{s-\ell}(D_x) + 1]\right]^{-1},\end{aligned}$$

so that we have the differential equation:

$$\left\{(1+x)\left[\prod_{\ell=0}^{r}[E_{s-\ell}(D_x) + 1]\right]^{-1} \right.$$
$$\left. + x\left[\prod_{\ell=r+1}^{s}[E_{s-\ell}(D_x) + 1]\right]^{-1}\right\} E_s(D_x)\,\ell_n^{[r,s]}(x) = n\,\ell_n^{[r,s]}(x).$$

5.2 Example

Let $G(t,x)=\exp[\Lambda_2(t)+x\,\Lambda_3(t)]=\sum_{n=1}^{\infty}\ell_n^{[2,3]}(x)\,\dfrac{t^n}{n!}$.

The first few $\ell_n^{[2,3]}(x)$ polynomials are:

$$\ell_0^{[2,3]}(x)=1$$

$$\begin{aligned}
\ell_1^{[2,3]}(x)&=x+1\\
\ell_2^{[2,3]}(x)&=x^2-2x-2\\
\ell_3^{[2,3]}(x)&=x^3-9x^2+8x+7\\
\ell_4^{[2,3]}(x)&=x^4-20x^3+92x^2-54x-35\\
\ell_5^{[2,3]}(x)&=x^5-35x^4+360x^3-1140x^2+551x+228\\
\ell_6^{[2,3]}(x)&=x^6-54x^5+970x^4-6850x^3+16951x^2-7615x-1834\\
\ell_7^{[2,3]}(x)&=x^7-77x^6+2128x^5-26145x^4-142821x^3-296457x^2+130686x+17582
\end{aligned}$$

Further values can be easily achieved by using Wolfram Alpha$^{©}$.

Differential Equation

$$\Big\{(1+x)\Big[[E_3(D_x)+1][E_2(D_x)+1][E_1(D_x)+1]\Big]^{-1} \\ +\;x\;[E_0(D_x)+1]^{-1}\Big\}\,E_3(D_x)\;e_n^{[1,2]}(x)=n\,e_n^{[1,2]}(x),$$

where the series expansions of the $E_k(\cdot)+1$ functions ($k=0,1,2,3$) are reported in Eq. (3.7).

5.3 The Case Log-Log I.2

Being $r>s$, recalling the operational rules (3.1), we find:

$$\begin{aligned}\hat{M} &= (1+x)\left[\prod_{\ell=0}^{s}[\Lambda_{\ell-1}(E_s(D_x))+1]\right]^{-1} + \left[\prod_{\ell=s+1}^{r}[\Lambda_{\ell-1}(E_s(D_x))+1]\right]^{-1} \\ &= (1+x)\left[\prod_{\ell=0}^{s}[E_{s-\ell}(D_x)+1]\right]^{-1} + \left[\prod_{\ell=s+1}^{r}[\Lambda_{\ell-s-2}(D_x)+1]\right]^{-1},\end{aligned}$$

so that we have the differential equation:

$$\left\{(1+x)\left[\prod_{\ell=0}^{s}[E_{s-\ell}(D_x)+1]\right]^{-1} \right.$$
$$\left. + \left[\prod_{\ell=s+1}^{r}[\Lambda_{\ell-s-2}(D_x)+1]\right]^{-1}\right\} E_s(D_x)\,\tilde{\ell}_n^{[r,s]}(x) = n\,\tilde{\ell}_n^{[r,s]}(x).$$

5.4 Example

Let $G(t,x) = \exp[\Lambda_3(t) + x\,\Lambda_2(t)] = \sum_{n=1}^{\infty} \tilde{\ell}_n^{[3,2]}(x)\,\frac{t^n}{n!}$.

The first few $\tilde{\ell}_n^{[3,2]}(x)$ polynomials are:

$$\tilde{\ell}_0^{[3,2]}(x) = 1$$

$$\tilde{\ell}_1^{[3,2]}(x) = x+1$$

$$\tilde{\ell}_2^{[3,2]}(x) = x^2 - x - 3$$

$$\tilde{\ell}_3^{[3,2]}(x) = x^3 - 6x^2 - 3x + 15$$

$$\tilde{\ell}_4^{[3,2]}(x) = x^4 - 14x^3 + 33x^2 + 69x - 105$$

$$\tilde{\ell}_5^{[3,2]}(x) = x^5 - 25x^4 + 165x^3 - 120x^2 - 1003x + 947$$

$$\tilde{\ell}_6^{[3,2]}(x) = x^6 - 39x^5 + 480x^4 - 1860x^3 - 1383x^2 + 14842x - 10472$$

$$\tilde{\ell}_7^{[3,2]}(x) = x^7 - 56x^6 + 1092x^5 - 8610x^4 + 19047x^3 + 60571x^2 - 238843x + 137337$$

Further values can be easily achieved by using Wolfram Alpha©.

Differential Equation

$$\Big\{(1+x)\Big[[E_2(D_x)+1][E_1(D_x)+1][E_0(D_x)+1]\Big]^{-1} + x\,[D_x+1]^{-1}\Big\}\,E_2(D_x)\,\tilde{\ell}_n^{[3,2]}(x) = n\,\tilde{\ell}_n^{[3,2]}(x),$$

where the series expansions of the $E_k(\cdot)+1$ functions $(k=0,1,2)$ and the $\Lambda_1(\cdot)$ function are reported in Eqs. (3.7) and (3.8).

6 The Case Exp-Log

We have:

$$A(t) = \exp[E_r(t)], \qquad H(t) = \Lambda_s(t),$$

$$\frac{A'(t)}{A(t)} = \prod_{\ell=0}^{r}[E_\ell(t)+1], \qquad H'(t) = \left[\prod_{\ell=0}^{s}[\Lambda_{\ell-1}(t)+1]\right]^{-1},$$

$$H^{-1}(t) = E_s(t) = f(t),$$

so that

$$\hat{P} = E_s(D_x),$$

$$\hat{M} = \prod_{\ell=0}^{r}[E_\ell(E_s(D_x))+1] + x\left[\prod_{\ell=0}^{s}[\Lambda_{\ell-1}(E_s(D_x))+1]\right]^{-1}.$$

Recalling the operational rules (3.1), we find:

$$\begin{aligned}\hat{M} &= \prod_{\ell=0}^{r}[E_\ell(E_s(D_x))+1] + x\left[\prod_{\ell=0}^{s}[\Lambda_{\ell-1}(E_s(D_x))+1]\right]^{-1}\\ &= \prod_{\ell=0}^{r}[E_{\ell+s+1}(D_x)+1] + x\left[\prod_{\ell=0}^{s}[E_{s-\ell}(D_x)+1]\right]^{-1},\end{aligned}$$

so that we have the differential equation:

$$\left\{\prod_{\ell=0}^{r}[E_{\ell+s+1}(D_x)+1] + x\left[\prod_{\ell=0}^{s}[E_{s-\ell}(D_x)+1]\right]^{-1}\right\} E_s(D_x)\,\varepsilon_n^{[r,s]}(x) = n\,\varepsilon_n^{[r,s]}(x).$$

6.1 Example

Let $G(t,x) = \exp[E_3(t) + x\,\Lambda_2(t)] = \sum_{n=1}^{\infty} \varepsilon_n^{[3,2]}(x)\,\dfrac{t^n}{n!}$.

The first few $\varepsilon_n^{[3,2]}(x)$ polynomials are:

$$\varepsilon_0^{[3,2]}(x) = 1$$

$$\begin{aligned}
\varepsilon_1^{[3,2]}(x) &= x+1\\
\varepsilon_2^{[3,2]}(x) &= x^2 - x + 5\\
\varepsilon_3^{[3,2]}(x) &= x^3 - 6x^2 + 21x + 35\\
\varepsilon_4^{[3,2]}(x) &= x^4 - 14x^3 + 81x^2 + 5x + 315\\
\varepsilon_5^{[3,2]}(x) &= x^5 - 25x^4 + 245x^3 - 640x^2 + 1697x + 3455\\
\varepsilon_6^{[3,2]}(x) &= x^6 - 39x^5 + 600x^4 - 3620x^3 + 11757x^2 + 4390x + 44590\\
\varepsilon_7^{[3,2]}(x) &= x^7 - 56x^6 + 1260x^5 - 12950x^4 + 69027x^3 - 121961x^2 + 294683x + 660665
\end{aligned}$$

Further values can be easily achieved by using Wolfram Alpha©.

Remark 6.1. Note that the sequence $\{1, 1, 5, 35, 315, 3455, 44590, 660665, \dots\}$ appears in the Encyclopedia of Integer Sequences [27] under A000357—Number of 5-level labeled rooted trees with n leaves.

Differential Equation

$$\Big\{[E_3(D_x)+1][E_4(D_x)+1][E_5(D_x)+1][E_6(D_x)+1]$$
$$+\,x\Big[[E_0(D_x)+1][E_1(D_x)+1][E_2(D_x)+1]\Big]^{-1}\Big\}\,E_2(D_x)\,\varepsilon_n^{[3,2]}(x) = n\,\varepsilon_n^{[3,2]}(x),$$

where $E_0(D_x) + 1 = \exp(D_x)$ and the series expansions of the $E_k(\cdot) + 1$ functions $(k = 1, 2, 3, 4, 5, 6)$ are reported in Eq. (3.7).

It is convenient to write the above equation in this form:

$$\Big\{\Big[\prod_{h=0}^{6}[E_h(D_x)+1]\Big][E_2(D_x)+1] - \Big[\prod_{h=0}^{6}[E_h(D_x)+1]\Big]\Big\}\,\varepsilon_n^{[3,2]}(x)$$
$$+\,x\,E_2(D_x)\,\varepsilon_n^{[3,2]}(x) = n\Big[\prod_{h=0}^{2}[E_h(D_x)+1]\,\varepsilon_n^{[3,2]}(x).$$

Then, we can apply the Cauchy multi-product formula. For instance, we have:

$$\Big[\prod_{h=0}^{6}[E_h(D_x)+1]\Big][E_2(D_x)+1]$$

$$=\sum_{\substack{k_0+k_1+\cdots+k_7=k\\ 0\le k\le\infty}}\binom{k}{k_0,k_1,\ldots,k_7}\prod_{\ell=0}^{6}b_{k_\ell}^{[\ell]}\cdot b_{k_7}^{[2]}\,\frac{D_x^k}{k!},$$

and in similar way the other multi-products follow.

7 The Case Log-Exp

We have:

$$A(t)=\exp[\Lambda_r(t)],\qquad H(t)=E_s(t),$$

$$\frac{A'(t)}{A(t)}=\left[\prod_{\ell=0}^{r}[\Lambda_{\ell-1}(t)+1]\right]^{-1},\qquad H'(t)=\prod_{\ell=0}^{s}[E_\ell(t)+1],$$

$$H^{-1}(t)=\Lambda_s(t)=f(t),$$

so that

$$\hat{P}=\Lambda_s(D_x),$$

$$\hat{M}=\left[\prod_{\ell=0}^{r}[\Lambda_{\ell-1}(\Lambda_s(D_x))+1]\right]^{-1}+x\prod_{\ell=0}^{s}[E_\ell(\Lambda_s(D_x))+1].$$

Recalling the operational rules (3.1), we find:

$$\hat{M}=\left[\prod_{\ell=0}^{r}[\Lambda_{\ell-1}(\Lambda_s(D_x))+1]\right]^{-1}+x\prod_{\ell=0}^{s}[E_\ell(\Lambda_s(D_x))+1]$$

$$=\left[\prod_{\ell=0}^{r}[\Lambda_{\ell+s}(D_x)+1]\right]^{-1}+x\prod_{\ell=0}^{s}[\Lambda_{s-\ell-1}(D_x)+1],$$

so that we have the differential equation:

$$\left\{\left[\prod_{\ell=0}^{r}[\Lambda_{\ell+s}(D_x)+1]\right]^{-1}+x\prod_{\ell=0}^{s}[\Lambda_{s-\ell-1}(D_x)+1]\right\}\Lambda_s(D_x)\,\lambda_n^{[r,s]}(x)=n\,\lambda_n^{[r,s]}(x).$$

7.1 Example

Let $G(t,x)=\exp[\Lambda_1(t)+x\,E_3(t)]=\sum_{n=1}^{\infty}\lambda_n^{[1,3]}(x)\,\frac{t^n}{n!}$.

The first few $\lambda_n^{[1,3]}(x)$ polynomials are:

$$\lambda_0^{[1,3]}(x)=1$$

$$\lambda_1^{[1,3]}(x)=x+1$$

$$\lambda_2^{[1,3]}(x)=x^2+6x-1$$

$$\lambda_3^{[1,3]}(x)=x^3+15x^2+31x+2$$

$$\lambda_4^{[1,3]}(x)=x^4+28x^3+178x^2+226x-6$$

$$\lambda_5^{[1,3]}(x)=x^5+45x^4+570x^3+2230x^2+1904x+24$$

$$\lambda_6^{[1,3]}(x)=x^6+66x^5+1385x^4+10990x^3+30154x^2+19093x-120$$

$$\lambda_7^{[1,3]}(x)=x^7+91x^6+2849x^5+37940x^4+214564x^3+444703x^2+216472x+720$$

Further values can be easily achieved by using Wolfram Alpha©.

The (absolute value) of the above polynomials, computed at $x=0$, are contained in the Encyclopedia of integer sequences [27], under A000142.

Differential Equation

$$\left\{\left[[\Lambda_3(D_x)+1][\Lambda_4(D_x)+1]\right]^{-1}\right.$$
$$\left.+\,x\,[\Lambda_2(D_x)+1][\Lambda_1(D_x)+1][\Lambda_0(D_x)+1][D_x+1]\right\}\Lambda_3(D_x)\,\lambda_n^{[1,3]}(x)=n\,\lambda_n^{[1,3]}(x),$$

where the series expansions of the $\Lambda_k(\cdot)+1$ functions $(k=0,1,2,3,4)$ function are reported in equations Remark 3.3, and by Eq. (3.2), $\Lambda_{-1}(D_x)+1:=D_x+1$. It is convenient to write the above equation in this form:

$$\left\{\Lambda_3(D_x) + x \prod_{h=-1}^{4} [\Lambda_h(D_x) + 1]\, \Lambda_3(D_x)\right\} \lambda_n^{[1,3]}(x) = n \prod_{h=3}^{4}[\Lambda_h(D_x) + 1]\, \lambda_n^{[1,3]}(x).$$

Then, we can apply the Cauchy multi-product formula. For instance, we have:

$$(D_x + 1) \prod_{n=0}^{4}[\Lambda_n(D_x) + 1]$$

$$= (D_x + 1) \sum_{\substack{k_0+k_1+\cdots+k_4=k \\ 0\le k\le\infty}} \binom{k}{k_0, k_1, \ldots, k_4} \prod_{\ell=0}^{4} l_{k_\ell}^{[\ell]} \frac{D_x^k}{k!},$$

and in similar way the other product follows.

8 Conclusion

We have introduced general sets of Exponential-Bell polynomials and their Logarithmic counterparts. The resulting polynomials satisfy operational differential equations whose coefficients are expressed in terms of generalized Bell or logarithmic numbers. The resulting connections of the values at the origin—or at the point $x = 1$—of some polynomial sets with the generalized Bell and logarithmic numbers have been noticed.

Acknowledgements The third author gratefully acknowledge the support of the organizers, who made possible his participation into ICDDEA 2019.

Author Contributions The authors claim to have contributed equally and significantly in this paper. Both authors read and approved the final manuscript.

Compliance with Ethical Standards
Conflict of Interest The authors declare that they have not received funds from any institution and that they have no conflict of interest.

References

1. Bell, E.T.: Exponential polynomials. Ann. Math. **35**, 258–277 (1934)
2. Bell, E.T.: The iterated exponential integers. Ann. Math. **39**, 539–557 (1938)
3. Ben Cheikh, Y.: Some results on quasi-monomiality. Appl. Math. Comput. **141**, 63–76 (2003)
4. Bernardini, A., Natalini, P., Ricci, P.E.: Multi-dimensional Bell polynomials of higher order. Comput. Math. Appl. **50**, 1697–1708 (2005)
5. Boas, R.P., Buck, R.C.: Polynomials defined by generating relations. Am. Math. Mon. **63**, 626–632 (1958)

6. Boas, R.P., Buck, R.C.: Polynomial Expansions of Analytic Functions. Springer, Berlin, Gottingen, Heidelberg, New York (1958)
7. Brenke, W.C.: On generating functions of polynomial systems. Am. Math. Mon. **52**, 297–301 (1945)
8. Bretti, G., Natalini, P., Ricci, P.E.: A new set of Sheffer-Bell polynomials and logarithmic numbers. Georgian Math. J. (2018). https://doi.org/10.1515/gmj-2019-2007
9. Bretti, G., Natalini, P., Ricci, P.E.: New sets of Euler-type polynomials. J. Ana. Num. Theor. **6**(2), 51–54 (2018)
10. Dattoli, G.: Hermite-Bessel and Laguerre-Bessel functions: a by-product of the monomiality principle. In: Cocolicchio, D., Dattoli, G., Srivastava, H.M. (eds.) Advanced Special Functions and Applications. Proceedings of the Melfi School on Advanced Topics in Mathematics and Physics; Melfi, Aracne Editrice, Rome, 9–12 May 1999, pp. 147–164 (2000)
11. Dattoli, G., Ricci, P.E., Srivastava, H.M.: Advanced special functions and related topics in probability and in differential equations. In: Proceedings of the Melfi School on Advanced Topics in Mathematics and Physics; Melfi, 24–29 June 2001. Appl. Math. Comput. **141**(1), 1–230 (2003)
12. Dattoli, G., Germano, B., Martinelli, M.R., Ricci, P.E.: Monomiality and partial differential equations. Math. Comput. Model. **50**, 1332–1337 (2009)
13. Natalini, P., Ricci, P.E.: An extension of the Bell polynomials. Comput. Math. Appl. **47**, 719–725 (2004)
14. Natalini, P., Ricci, P.E.: Higher order Bell polynomials and the relevant integer sequences. Appl. Anal. Discrete Math. **11**, 327–339 (2017)
15. Natalini, P., Ricci, P.E.: Remarks on Bell and higher order Bell polynomials and numbers. Cogent Math. **3**, 1–15 (2016)
16. Natalini, P., Ricci, P.E.: New Bell-Sheffer polynomial sets. Axioms **7**, 71 (2018). https://doi.org/10.3390/axioms7040071
17. Natalini, P., Ricci, P.E.: Bell-Sheffer exponential polynomials of the second kind. Georgian Math. J. (2018, in press)
18. Qi, F., Niu, D.-W., Lim, D., Guo, B.-N.: Some properties and an application of multivariate exponential polynomials. HAL Archives (2018). https://hal.archives-ouvertes.fr/hal-01745173
19. Qi, F.: Integral representations for multivariate logarithmic potentials. J. Comput. Appl. Math.**336**, 54–62 (2018). https://doi.org/10.1016/j.cam.2017.11.047
20. Qi, F.: On multivariate logarithmic polynomials and their properties. Indag. Math. **336** (2018, in press). https://doi.org/10.1016/j.indag.2018.04.002
21. Ricci, P.E., Natalini, P., Bretti, G.: Sheffer and Brenke polynomials associated with generalized Bell numbers. J$\tilde{n}\bar{a}n\bar{a}$bha **47**(2), 337–352 (2017)
22. Ricci, P.E.: Logarithmic-Sheffer polynomial sets. J$\tilde{n}\bar{a}n\bar{a}$bha **48**(1), 119–128 (2018)
23. Ricci, P.E.: Logarithmic-Sheffer polynomials of the second kind. Tbilisi Math. J. **11**(3), 95–106 (2018)
24. Ricci, P.E.: Complex spirals and pseudo-Chebyshev polynomials of fractional degree. Axioms **7**, 71 (2018). https://doi.org/10.3390/axioms7040071
25. Roman, S.M.: The Umbral Calculus. Academic Press, New York (1984)
26. Sheffer, I.M.: Some properties of polynomials sets of zero type. Duke Math. J. **5**, 590–622 (1939)
27. Sloane, N.J.A.: The On-Line Encyclopedia of Integer Sequences (2016). http://oeis.org
28. Srivastava, H.M., Manocha, H.L.: A Treatise on Generating Functions. Halsted Press (Ellis Horwood Limited, Chichester); Wiley, New York, Chichester, Brisbane and Toronto (1984)
29. Steffensen, J.F.: The poweroid, an extension of the mathematical notion of power. Acta Math. **73**, 333–366 (1941)

Existence Results for Periodic Boundary Value Problems with a Convection Term

Pasquale Candito and Roberto Livrea

Abstract By using an abstract coincidence point theorem for sequentially weakly continuous maps the existence of at least one positive solution is obtained for a periodic second order boundary value problem with a reaction term involving the derivative u' of the solution u: the so called convention term. As a consequence of the main result also the existence of at least one positive solution is obtained for a parameter-depending problem.

2010 AMS Subject Classification: 34B15 · 34B18

1 Introduction

The aim of this paper is to obtain new existence results for the following periodic boundary value problem

$$\begin{cases} -u'' + M(t)u = f(t, u, u') & \text{in } (0, T) \\ u(T) - u(0) = u'(T) - u'(0) = 0, \end{cases} \tag{1}$$

where $T > 0$, $M : [0, T] \to \mathbf{R}$ is a continuous and positive function and $f : [0, T] \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is a continuous function with $f(t, 0, 0) \neq 0$, for every $t \in [0, T]$.

The research was supported by the Ministry of Education, University and Research of Italy, Prin 2017 "Nonlinear Differential Problems via Variational, Topological and Set-valued Methods", Project No. 2017AYM8XW.

P. Candito (✉)
Department DICEAM, University of Reggio Calabria,
Via Graziella - Feo di Vito, 89100 Reggio Calabria, Italy
e-mail: pasquale.candito@unirc.it

R. Livrea
Department of Mathematics and Computer Science, University of Palermo,
Via Archirafi, 34, 90123 Palermo, Italy
e-mail: roberto.livrea@unipa.it

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_43

As usual, here we say that problem (1) has a convention term because the nonlinearity f depends both on the function u and its derivative u'.

Concerning boundary value problems there is a well consolidated literature where many pioneering results are obtained by several scholars using different tools, as for instance, a priori bounds and topological degree [8, 10, 22]; upper and lower methods [7, 14, 24] and fixed point theory [1, 11].

In particular, as pointed out in [25], the application of the fixed point theorem in studying problem (1) is strictly connected to the sign properties of the Green's function associated to the linear homogeneous problem, that is $f \equiv 0$.

Recently, many authors paid attention to this topic and very interesting results are pointed out in [2, 5, 12, 13, 15, 18, 20, 26, 27].

Here, for obtaining our main results, we apply a coincidence point theorem for sequentially weakly continuous maps [3], see Theorem 1 below, in the variational setting used in [23]. Such approach in spirit is based on an useful version of Fan's fixed point theorem [9] contained in [4]. However, we do not use the Green's function to get the solutions of problem (1). Moreover, we do not require any asymptotic growth condition on the nonlinearity f at zero and/or at infinity. We just assume condition (9) below, together $f(t, 0, 0) \neq 0$, for every $t \in [0, T]$ to guarantee the existence of a nontrivial solution which become positive provided that $f(t, 0, 0) > 0$ for every $t \in [0, T]$.

However, as far as we know, there are few papers dealing with problem (1). For example, in [19], applying a coincidence degree theorem and when the nonlinear term is of the form $f(t, x, y) = h(t)g(x, y)$, the existence of at least one positive solution is ensured in terms of the relative behaviors of $\frac{g(x,y)}{|x|+|y|}$ for $|x| + |y|$ near 0 and $+\infty$, where

(H) $h : [0, T] \to [0, +\infty)$ and $g : [0, +\infty) \times \mathbf{R} \to [0, +\infty)$ are continuous, $h(t) \not\equiv 0$.

Furthermore, for the readers interested to the applications of periodic BVP in physics and engineering, we again mention [19] and the references therein.

On the other hand, it seems that much more attention is paid to problems without convention terms and depending from a positive parameter λ. An example is the following

$$\begin{cases} -u'' + M(t)u = \lambda g(t, u) & \text{in } (0, T) \\ u(T) - u(0) = u'(T) - u'(0) = 0, \end{cases} \tag{2}$$

where $T > 0$, $M : [0, T] \to \mathbf{R}$ is a continuous and positive function and $g : [0, T] \times \mathbf{R} \to \mathbf{R}$ is a continuous function.

In this case, many existence, non-existence and multiplicity results have been obtained, for instance, in [12, 13, 16, 17, 20, 21, 27], requiring suitable asymptotic behaviors of the "slope "$f(t, u)/u$ of f at zero and at infinity.

Finally, for the sake of completeness, we wish to stress that in [3] and [6] a similar approach to those proposed in the present note has been adopted for the study of a Dirichlet and a Neumann boundary value problem respectively.

2 Preliminaries

We recall that the weak derivative of a function $u \in L^1([0, T])$ is a function $u' \in L^1([0, T])$ such that

$$\int_0^T u(t)\varphi'(t)\,dt = -\int_0^T u'(t)\varphi(t)\,dt$$

for every $\varphi \in C_T^\infty$, where C_T^∞ is the space of indefinitely differentiable T-periodic functions (see [23]).

Let us denote by H_T the Sobolev space of functions $u \in L^2([0, T])$ having a weak derivative $u' \in L^2([0, T])$, while

$$H_T^2 = \{u \in H_T : u' \in H_T\}.$$

According to ([23, pp. 6–7]), for every $u \in H_T^2$ one has that

$$\int_0^T u'(t)\,dt = \int_0^T u''(t)\,dt = 0,$$

hence the periodic conditions $u(T) - u(0) = u'(T) - u'(0) = 0$ hold. Moreover, if we endow H_T^2 with the norm

$$\|u\| = \|u\|_2 + \|u'\|_2 + \|u''\|_2$$

for every $u \in H_T^2$ and on $C^1([0, T])$ we consider the norm

$$\|u\|_{C^1} = \max\{\|u\|_\infty, \|u'\|_\infty\},$$

H_T^2 is compactly embedded in $C^1([0, T])$, see [23, Proposition 1.2]. In particular, if $u \in H_T^2$ observe that

$$\begin{aligned} |u(t)| &= \frac{1}{T}\left|\int_0^T u(s) + \int_0^T \left(\int_s^t u'(x)\,dx\right) ds\right| \\ &\leq \frac{1}{T}\|u\|_1 + \|u'\|_1 \leq T^{-1/2}\|u\|_2 + T^{1/2}\|u'\|_2 \\ &\leq \max\{T^{-1/2},\ T^{1/2}\}\|u\| \end{aligned}$$

for every $t \in [0, T]$. Thus, if we put

$$c_T = \max\{T^{-1/2},\ T^{1/2}\}, \tag{3}$$

one can conclude that

$$\|u\|_\infty \leq c_T\|u\|. \tag{4}$$

Similarly one can obtain

$$\|u'\|_\infty \le c_T \|u\|, \tag{5}$$

namely

$$\|u\|_{C^1} \le c_T \|u\|. \tag{6}$$

Incidentally, observe that if $0 < T \le 1$ then $c_T = T^{-1/2}$ and one can realize the equality in (6) choosing u constant. Namely, if $0 < T \le 1$ the constant introduced in (3) is the best one of the embedding. Some sharp estimates for the norms of functions in H_T can be found in [23, Proposition 1.3].

A direct computation based on (6) shows that for every $r > 0$

$$B_r = \{u \in H_T^2 : \|u\| \le r\} \subseteq \{u \in C^1([0,1]) : \ \|u\|_{C^1} \le c_T r\}. \tag{7}$$

The following coincidence point theorem represents the key tool for the proof of our main results.

Theorem 1. *Let X, Y be real Banach spaces, let K be a weakly compact, convex subset of X, and let F, G be sequentially weakly continuous functions from K into Y, that is, if $x_n \rightharpoonup x$ in K then $F(x_n) \rightharpoonup F(x)$ and $G(x_n) \rightharpoonup G(x)$ in Y. Assume that $F^{-1}(y)$ is a nonempty convex set for all $y \in G(K)$. Then there exists $x_0 \in K$ such that $F(x_0) = G(x_0)$.*

3 Main Results

Here is the first existence result for the considered periodic problem.

Theorem 2. *Let $f : [0,T] \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be a continuous function. Put*

$$\tau = \frac{\mu}{c_T\sqrt{T}\,[1 + (T+1)(\|M\|_\infty + \mu)]}, \tag{8}$$

with $\mu = \min_{t\in[0,T]} M(t)$, and assume that there exists $r > 0$ such that

$$\max_{(t,x,y)\in[0,T]\times[-r,r]\times[-r,r]} |f(t,x,y)| \le \tau \cdot r. \tag{9}$$

Then, problem (1) admits at least one classical solution $\tilde{u}$ such that

$$(\tilde{u}(t), \tilde{u}'(t), \tilde{u}''(t)) \in \ [-r,r] \times [-r,r] \times [-(\|M\|_\infty + \tau)r, (\|M\|_\infty + \tau)r].$$

Proof. We will apply Theorem 1 with $X = H^2_T$, $Y = X^*$, $K = B_\rho$, being $\rho = \frac{r}{c_T}$, and $F, G : X \to X^*$ the functions defined as follows

$$F(u)(v) = \int_0^T (u'(t)v'(t) + M(t)u(t)v(t))\, dt,$$

$$G(u)(v) = \int_0^T f(t, u(t), u'(t))\, dt$$

for every $u,\ v \in X$. Indeed, K is weakly compact in view of the reflexivity of X, while the compactness of the embedding of X into $C^1([0, T])$ assures that both F and G are sequentially weakly continuous functions from X to X^*.
We claim that

$$G(K) \subseteq F(K). \tag{10}$$

Fix $w^* \in G(K)$ and let $w \in K$ be such that $G(w) = w^*$. Put

$$g(t) = f(t, w(t), w'(t))$$

for all $t \in [0, T]$ and observe that $g \in C^0([0, T])$. Hence, applying the Minty-Browder theorem (or the Lax-Milgram theorem) in the space H_T, the following problem

$$\begin{cases} -u'' + M(t)u = g(t) & \text{in } (0, T) \\ \\ u(T) - u(0) = u(T) - u(0) = 0 \end{cases} \tag{11}$$

admits a unique weak solution $u_w \in H_T$ and, in particular, thanks to the classical regularity theory, one has that $u_w \in C^2([0, T])$ and it is a classical solution.
If we localize $u_w \in H^2_T$ and prove that

$$u_w \in B_\rho, \tag{12}$$

we can conclude that (10) holds, since $F(u_w) = G(w) = w^*$.
To this end, we first point out that

$$\|u_w\|_\infty \le \frac{\|g\|_\infty}{\mu}, \tag{13}$$

$$\|u'_w\|_\infty \le T\left(\frac{\|M\|_\infty}{\mu} + 1\right)\|g\|_\infty, \tag{14}$$

and

$$\|u''_w\|_\infty \le \left(\frac{\|M\|_\infty}{\mu} + 1\right)\|g\|_\infty. \tag{15}$$

Indeed, fix $k = \frac{\|g\|_\infty}{\mu}$ and put $\varphi(t) = (u_w - k)^+$. Obviously $\varphi \in H_T$ and $\varphi' = u'_w \cdot \chi_{\{u_w \geq k\}}$. Hence, from (11) one has

$$\int_0^T (u'_w \varphi' + M(t) u_w \varphi)\, dt = \int_0^T g\varphi\, dt$$

that is

$$\begin{aligned} 0 &\leq \int_0^T M(t)(u_w - k)(u_w - k)^+\, dt \\ &\leq \int_0^T ((u'_w)^2 \chi_{\{u_w \geq k\}} + M(t)(u_w - k)(u_w - k)^+)\, dt \\ &= \int_0^T (g - M(t)k)(u_w - k)^+\, dt \leq 0, \end{aligned}$$

and this implies that $(u_w - k)(u_w - k)^+ \equiv 0$, namely

$$u_w(t) \leq k \tag{16}$$

for every $t \in [0, T]$. Arguing in a similar way, one has that

$$-k \leq u_w(t) \tag{17}$$

for every $t \in [0, T]$. Clearly (16) and (17) lead to (13).
Moreover, since $u_w(0) = u_w(T)$, there exists $t_0 \in (0, T)$ such that $u'_w(t_0) = 0$ and, in view of (13), for every $t \in [0, T]$ one has

$$\begin{aligned} |u'_w(t)| &= \left| \int_{t_0}^t u''_w(s)\, ds \right| \\ &= \left| \int_{t_0}^t (M(s)u_w(s) - g(s))\, ds \right| \\ &\leq T(\|M\|_\infty \|u_w\|_\infty + \|g\|_\infty) \\ &\leq T\left(\frac{\|M\|_\infty}{\mu} + 1 \right) \|g\|_\infty, \end{aligned}$$

namely (14) holds.
Exploiting again that u_w is a classical solution of problem (11), from (13) one derives

$$\|u''_w\|_\infty \leq \left(\frac{\|M\|_\infty}{\mu} + 1 \right) \|g\|_\infty$$

and (15) is verified.

Now observe that from (7) it follows that $\|w\|_{C^1} \leq r$, hence, in view of assumption (9), $\|g\|_\infty \leq \tau \cdot r$. Putting together (13)–(15) and this last estimate, one has

$$\|u_w\|_2 + \|u_w'\|_2 + \|u_w''\|_2 \leq \tau \frac{\sqrt{T}}{\mu} \left[1 + (T+1)(\|M\|_\infty + \mu)\right] r = \frac{r}{c_T} = \rho,$$

namely (12) holds and (10) is verified.
It is simple to verify that F is injective, hence $F^{-1}(w^*) = \{u_w\}$ for every $w^* \in G(K)$ and all the assumptions of Theorem 1 are satisfied. Thus, there exists $\tilde{u} \in K$ such that

$$F(\tilde{u})(v) = G(\tilde{u})(v)$$

for every $v \in H_T^2$. But $C_T^\infty \subset H_T^2$ implies that $\tilde{u}' \in H_T$, being $M(t)\tilde{u} - f(t, \tilde{u}, \tilde{u}')$ its weak derivative. The regularity theory assures that $\tilde{u} \in C^2([0,T])$ and it is a classical solution of (1). The proof is complete since $\|\tilde{u}\|_\infty$, and $\|\tilde{u}'\|_\infty$ can be estimated recalling (7), while $\|\tilde{u}''\|_\infty$ can be estimated exploiting the fact that $\tilde{u}$ solves (1).

As a consequence of the previous result, we can state the main constant sign periodic solution theorem.

Theorem 3. *Let $f : [0,T] \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be a continuous function such that $f(t,0,0) > 0$ for every $t \in [0,T]$. Let $\tau > 0$ as defined in (8) and assume that*

$$\max_{(t,x,y)\in[0,T]\times[0,r]\times[-r,r]} |f(t,x,y)| \leq \tau \cdot r. \tag{18}$$

Then, problem (1) admits at least one positive classical solution $\tilde{u}$ such that

$$(\tilde{u}(t), \tilde{u}'(t), \tilde{u}''(t)) \in \ (0,r] \times (0,r] \times [-(\|M\|_\infty + \tau)r, (\|M\|_\infty + \tau)r].$$

Proof. We make use of some truncation arguments. Let $\hat{f} : [0,T] \times \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be the function defined by

$$\hat{f}(t,x,y) = \begin{cases} f(t,x,y) & \text{if } x \geq 0 \\ f(t,0,y) & \text{if } x < 0. \end{cases} \tag{19}$$

If we consider the following auxiliary periodic problem

$$\begin{cases} -u'' + M(t)u = \hat{f}(t,u,u') & \text{in } [0,T] \\ u(T) - u(0) = u'(T) - u(0) = 0, \end{cases} \tag{20}$$

it is evident that the non negative solutions of (20) are also constant sign solutions of problem (1). At this point, we can observe that, thanks to (18) and (19), $\hat{f}$ satisfies all the assumptions of Theorem 2. Hence, problem (20) admits at least one classical solution $\tilde{u} \in C^2([0,T])$. Finally, the proof is complete if we verify that

$$\min_{t\in[0,T]} \tilde{u}(t) > 0. \tag{21}$$

Suppose (21) false, namely, there exists $t^* \in [0, T]$ such that

$$\tilde{u}(t^*) = \min_{t \in [0,T]} \tilde{u}(t) \leq 0.$$

Thus, we have that

$$\tilde{u}'(t^*) = 0, \quad \tilde{u}''(t^*) \geq 0. \tag{22}$$

Indeed, if $t^* \in (0, T)$ then (22) is obvious. Otherwise, suppose that $t^* = 0$ (the other case $t^* = T$ is analogous). Since 0 is a minimizer of $\tilde{u}$ one has that $\tilde{u}'(0) \geq 0$, but the periodic boundary conditions lead to $\tilde{u}'(0) = 0$. Otherwise, if $\tilde{u}'(0) > 0$ one has $\tilde{u}'(T) > 0$ and for t close to T one achieves the contradiction $\tilde{u}(t) < \tilde{u}(T) = \tilde{u}(0) = \min_{[0,T]} \tilde{u}$.

Moreover, if it was $\tilde{u}''(0) < 0$, since $\tilde{u} \in C^2([0, T])$, one could find a suitable $\delta > 0$ such that $\tilde{u}'(t) < 0$ for all $t \in (0, \delta)$, in contradiction with the fact that $t^* = 0$ is a minimizer.

At this point, exploiting (22) one is lead to the evident contradiction

$$0 \geq -\tilde{u}''(t^*) + M(t^*)\tilde{u}(t^*) = \hat{f}(t^*, \tilde{u}(t^*), \tilde{u}'(t^*)) = f(t^*, 0, 0) > 0.$$

In conclusion, (21) holds and the proof is completed.

Remark 1. The existence of a negative classical solution can be similarly proved if one assumes that $f(t, 0, 0) < 0$ for every $t \in [0, T]$, in place of $f(t, 0, 0) > 0$.

Corollary 1. *Let $M : [0, T] \to \mathbf{R}$ be a continuous and positive function and $g : [0, T] \times \mathbf{R} \to \mathbf{R}$ a continuous function. Then, there exists $\lambda^* > 0$ such that, for each $\lambda \in]-\lambda^*, \lambda^*[$, problem (2) admits at least one classical solution.*

Proof. Let τ be as given in (8) and put

$$\lambda^* = \tau \sup_{r>0} \frac{r}{\max\limits_{[0,T]\times[-r,r]} |g(t, x)|}.$$

Therefore, fixed λ such that $|\lambda| < \lambda^*$, it is clear that there exists $r > 0$ such that

$$\max_{(t,x)\in[0,T]\times[-r,r]} |\lambda g(t, x)| < \tau r.$$

In few words, the function λg fulfils condition (9) of Theorem 2 and our conclusion follows.

Example 1. The following problem

$$\begin{cases} -u'' + \frac{u}{2} = \frac{2+\sin(t)}{40\pi^2}(1 - u^3)(1 - u'^4) & \text{in } [0, 2\pi] \\ u(2\pi) - u(0) = u'(2\pi) - u'(0) = 0, \end{cases} \tag{23}$$

admits at least one positive and non constant solution.

Indeed, we can apply Theorem 3 if we consider $r = 1$, $M(t) \equiv 1/2$ and put

$$f(t, x, y) = \frac{2 + \sin(t)}{40\pi^2}(1 - x^3)(1 - y^4)$$

for every $(t, x, y) \in [0, 1] \times \mathbf{R} \times \mathbf{R}$. Direct computations show that

$$\begin{aligned}\max_{[0,1]\times[0,1]\times[-1,1]} |f(t, x, y)| &= \max_{[0,1]\times[0,1]\times[-1,1]} \frac{2 + \sin(t)}{40\pi^2}(1 - x^3)(1 - y^4) \\ &= \frac{3}{40\pi^2},\end{aligned}$$

namely (18) is satisfied, being $\tau = \frac{1}{8\pi(1+\pi)}$. Hence, (23) has at least one positive classical solution u_0 such that

$$(u_0(t), u_0'(t), u_0''(t)) \in (0, 1] \times (0, 1] \times \left[-\frac{1}{2} - \frac{1}{8\pi(1 + \pi)}, \frac{1}{2} + \frac{1}{8\pi(1 + \pi)}\right]$$

for every $t \in [0, 1]$. Finally, it is easy to verify that (23) does not admits constant solutions.

References

1. Amann, H.: Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces. SIAM Rev. **18**(4), 620–709 (1976)
2. Atici, F.M., Guseinov, G.S.: On the existence of positive solutions for nonlinear differential equations with periodic boundary conditions. J. Comput. Appl. Math. **132**, 341–356 (2001)
3. Bonanno, G., Candito, P., Motreanu, D.: A coincidence point theorem for sequentially continuous mappings. J. Math. Anal. Appl. **435**(1), 606–615 (2016)
4. Bonanno, G., Marano, S.A.: Positive solutions of elliptic equations with discontinuous nonlinearities. Topol. Methods Nonlinear Anal. **8**, 263–273 (1996)
5. Cabada, A., Cid, J.: Existence and multiplicity of solutions for a periodic Hill's equation with parametric dependence and singularities. Abstr. Appl. Anal. **2011**, 19 (2011). Art. ID 545264
6. Candito, P., Livrea, R.: An existence result for a Neumann problem. Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal. **22**, 481–488 (2015)
7. De Coster, C., Habets, P.: Upper and lower solutions in the theory of ODE boundary value problems: classical and recent results. In: Zanolin, F. (ed.), Nonlinear analysis and boundary value problems for ordinary differential equations (Udine), CISM Courses and Lectures, vol. 371, pp. 1–78. Springer, Vienna (1996)
8. Drábek, P.: Landesman-Lazer condition for nonlinear problems with jumping nonlinearities. J. Differential Equations **85**, 186–199 (1990)
9. Fan, K.: Fixed-point and minimax theorems in locally convex topological linear spaces. Proc. Natl. Acad. Sci. U.S.A. **38**, 121–126 (1952)
10. Gaines, R.E., Mawhin, J.: Coincidence degree, and nonlinear differential equations. In: LNM, vol. 568. Springer, Berlin (1977)
11. Guo, D., Lakshmikantham, V.: Nonlinear Problems in Abstract Cones. Academic Press, Orlando (1988)

12. Graef, J., Kong, L., Wang, H.: A periodic boundary value problem with vanishing Green' s function. Appl. Math. Lett. **21**(2), 176–180 (2008)
13. Graef, J., Kong, L., Wang, H.: Existence, multiplicity, and dependence on a parameter for a periodic boundary value problem. Differ. Equ. **245**, 1185–1197 (2008)
14. Habets, P., Sanchez, L.: Periodic solutions of some Liénard equations with singularities. Proc. Amer. Math. Soc. **109**(4), 1035–1044 (1990)
15. Hai, D.D.: Existence of positive solutions for periodic boundary value problem with sign-changing Green's function. Positivity **22**(5), 1269–1279 (2018)
16. Hao, X., Liu, L., Wu, Y.: Existence and multiplicity results for nonlinear periodic boundary value problems. Nonlinear Anal. **72**, 3635–3642 (2010)
17. He, Z., Ma, R., Xu, M.: Positive solutions for a class of semipositone periodic boundary value problems via bifurcation theory. Electron. J. Qual. Theory Differ. Equ. **2019**(29), 15 (2019)
18. Jiang, D., Chu, J., O'Regan, D., Agarwal, R.: Multiple positive solutions to superlinear periodic boundary value problems with repulsive singular forces. J. Math. Anal. Appl. **286**, 563–576 (2003)
19. Liu, J., Feng, H.: Positive solutions of periodic boundary value problems for second-order differential equations with the nonlinearity dependent on the derivative. J. Appl. Math. Comput. **49**(1–2), 343–355 (2015)
20. Ma, R.: Nonlinear periodic boundary value problems with sign-changing Green's function. Nonlinear Anal. **74**(5), 1714–1720 (2011)
21. Ma, R., Xu, J., Han, X.: Global structure of positive solutions for superlinear second-order periodic boundary value problems. Appl. Math. Comput. **218**, 5982–5988 (2012)
22. Mawhin, J.: Topological degree and boundary value problems for nonlinear differential equations. In: Furi, M., Zecca, P. (eds.) Topological Methods for Ordinary Differential Equations. Lecture Notes in Mathematics, vol. 1537, pp. 74–142. Springer, New York (1993)
23. Mawhin, J., Willem, M.: Critical Point Theory and Hamiltonian Systems. Springer, New York (1989)
24. Rachůnková, I., Tvrdý, M.: Nonlinear systems of differential inequalities and solvability of certain boundary value problems. J. Inequal. Appl. **6**(2), 199–226 (2001)
25. Torres, P.J.: Existence of one-signed periodic solutions of some second-order differential equations via a Krasnoselskii fixed point theorem. J. Differ. Equ. **190**(2), 643–662 (2003)
26. Wang, H.: On the number of positive solutions of nonlinear systems. J. Math. Anal. Appl. **281**, 287–306 (2003)
27. Webb, J.R.L.: Boundary value problems with vanishing Green's function. Commun. Appl. Anal. **13**(4), 587–595 (2009)

Numerical Solution of the Time Fractional Cable Equation

M. Luísa Morgado, Pedro M. Lima, and Mariana V. Mendes

Abstract The time fractional diffusion equation has attracted the attention of many researchers in the last years due to its many applications in different domains. In this article we are concerned with one of these models, the time fractional cable equation:

$$\frac{\partial^2 V(x,t)}{\partial x^2} - \frac{\partial^\alpha V(x,t)}{\partial x^\alpha} - V(x,t) = 0, \tag{1}$$

which describes the spatial and temporal dependence of transmembrane potential $V(x,t)$ along the axial direction of a cylindrical nerve cell segment. Here $\frac{\partial^\alpha V(x,t)}{\partial x^\alpha}$ is a Caputo-type derivative, with $0 < \alpha < 1$. We use a numerical scheme to solve Eq. (1), which is based on the $L1$-method and on a finite-difference scheme for the time and space discretization, respectively. In order to deal with the singularity at $t = 0$ we use non-uniform meshes. Numerical examples are presented which illustrate the efficiency of the method.

1 Introduction

In this article we are concerned with a fractional model in Neurophysiology, a scientific field that studies the functioning of the nervous system based on tools like electrophysiological recordings, voltage clamp, extracellular single-unit recording

M. L. Morgado · P. M. Lima (✉) · M. V. Mendes
CEMAT, Instituto Superior Técnico, University of Lisbon,
Av. Rovisco Pais, 1049 001 Lisboa, Portugal
e-mail: plima@math.ist.utl.pt

M. L. Morgado
e-mail: luisam@utad.pt

M. V. Mendes
e-mail: marianamendes.mvm@gmail.com

M. L. Morgado
Departamento de Matemática, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_44

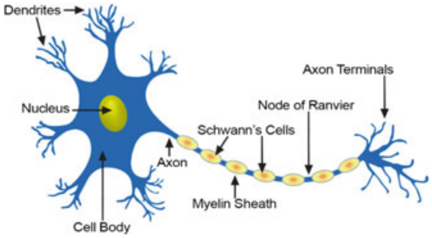


Fig. 1 Structure of a typical neuron (https://training.seer.cancer.gov/anatomy/nervous/tissue.html)

and recording of local field potentials. This area is related to many others, e.g. electrophysiology, neuroanatomy, mathematical neuroscience, and biophysics.

Our main objective is to model the electrical conduction in non-isopotential excitable cells called neurons. Before we start describing the process we need to look into the neuron structure that we can see in Fig. 1 .

Looking at the picture we can see that a dendrite looks like branches of a tree around the cell body of the neuron. It is through the dendrites that the neurons receive electrical signals.

1.1 Cable Equation in Neurophysiology

The cable equation has been used all over the years to study multiple applications. In this work we will consider an application to Neurophysiology.

The cable equation uses mathematical algorithms to calculate electric current along passive neurites (neuronal process that refers to any projection from the cell body of a neuron), particularly the dendrites that receive synaptic inputs at different sites and times. The neuron may be polarized by means of experimentally imposed voltages and currents. This behaviour of the neuron is well described by the one-dimensional cable equation.

In [5], the author explains the following: “The cable equation model of passive axons can thus be a valuable tool both for analyzing the means by which a neuron combines a pattern of synaptic inputs to produce a specific computation and for determining the functionally significant electrical parameters of the cell by measurements of the response to various applied currents. In order for the model to be of practical use, though, it must be possible first to set up the equation so that the boundary and initial conditions correspond to the anatomical, physiological, and experimental realities, and second, to solve the equation under these conditions.”

In the present work we are going to consider only the case of a finite cable, that is, we will consider a dendrite of finite length L which does not branch out or have any synaptic connections.

This is an important condition that we have to take into account when choosing a numerical scheme to approximate the solution.

1.2 The Time Fractional Cable Equation

As mentioned above, the cable equation has many applications, in particular, the one -dimensional cable model can be used in Neurophysiology. In this case it simulates electric current along neurites. As pointed out in [10], the equation "describes the spatial and temporal dependence of transmembrane potential $u(t, x)$ along the axial x direction of a cylindrical nerve cell segment".

The resulting differential equation for the transmembrane potential takes the form of a standard diffusion equation with an extra term to account leakage of ions out of the membrane, which results in a decay of the electric signal in space and in time.

We are dealing with a Cauchy Problem where we see the behaviour of the transmembrane potential when the system is excited at one end. We are going to apply the cable equation to a finite length cable. Therefore, we are considering the spacial rank between $x = 0$ and $x = L$ and time interval $[0, T]$. When applied to the problem under consideration, cable equation has the following form [2]:

$$r_m \frac{\partial u_m(t,x)}{\partial t} = \theta^2 \frac{\partial^2 u_m(t,x)}{\partial x^2} - u_m(t,x) + E_m + r_m I_{inj} I(t,x), \quad t \in [0,T], \quad x \in [0,L]. \tag{2}$$

In this equation $u_m(t, x)$ is the membrane potential in mV and $I_{inj}(t, x)$ is the injected current in Amperes. The remaining coefficients are constants: r_m is the membrane resistance per unit length of the fiber in Ωcm, $\nu = r_m c_m$ is the membrane time constant and $\theta = (r_m/r_a)^{0.5}$ is the membrane space constant. c_m is the capacitance per unit length of cable of diameter d in units of F/cm. r_a is the axial resistance per unit length in Ω/cm. E_m is the leakage reversal potential due to different ions in mV and it varies depending on the cell type, but we are going to consider $E_m = 0$ for the simplicity of the problem. In [10], instead of (2), the authors introduce an equation with non integer order temporal derivative:

$$\frac{\partial^\alpha U(\Upsilon, \chi)}{\partial \Upsilon^\alpha} = \frac{\partial^2 U(\Upsilon, \chi)}{\partial \chi^2} - U(\Upsilon, \chi) + \frac{I(\Upsilon, \chi)}{\theta c_m}, \quad t \in [0,T], \quad x \in [0,L] \tag{3}$$

where $I(\Upsilon, \chi) = \theta \nu I_{inj}(\Upsilon, \chi)$ and $I_{inj}(\Upsilon, \chi)$ is the applied stimulus current density also scaled (per unit length). $c_m = C_m \pi d$ and C_m is capacitance per unit area in F/cm^2. The diameter of the cable is d and it is in units of μm, θ is the membrane space constant. We also take $\Upsilon = t\nu$ and $\chi = x/\theta$.

The fractional cable model is more realistic than the standard integer-order cable equation. In fact, "the fractional cable model predicts that postsynaptic potentials propagating along dendrites with larger spine densities can arrive at the soma faster and be sustained at higher levels over longer times" (see [1]). Since we take $\alpha \in]0, 1[$, we expect that the solutions of the Eq. (3) describe better the qualitative behaviour of the membrane potential than the usual approach with $\alpha = 1$. Moreover, by changing α we can fit the numerical results to the experimental ones.

For the fractional derivative in time we are going to use the Caputo definition. The reason why we use this definition instead of any other is because it has already been proposed in [10] to model spiking adaptation for a homogeneous membrane patch. Since the Caputo fractional derivative is a non-local operator (see [8]), “it could be also introduced to explain behaviours like multiple timescale dynamics and memory effects, related to the complexity of the medium”, as pointed out in [10].

In order to make this equation realistic to the problem in study we need to establish the boundary and initial conditions in such a way that corresponds to the anatomical, physiological and experimental realities, and also to solve the equation under these conditions. The partial differential equation for a single unbranched cable has a unique solution only if boundary conditions are specified at the endpoints. Since we are going to consider the finite cable equation we define the terminations as $x = 0$ and $x = L$. There are several boundary conditions for the finite cable equation [2, 9] corresponding to different situations. For example, in the killed-end or Dirichlet case the voltage is clamped to zero and the axial current “leaks” out to ground. This is the case that we are going to use and it can be simulated using the L1 method, as we will describe in the next section. This is also the simplest case in which the end of the neurite has been cut. This can arise in some preparations such as dissociated cells, and it means that the intracellular and extracellular media are directly connected at the end of the neurite. Thus the membrane potential at the end of the neurite is equal to the extracellular potential.

$$u(0,t) = 0, \qquad t > 0 \tag{4}$$

$$u(L,t) = 0, \qquad t > 0. \tag{5}$$

The intracellular fluid therefore ends abruptly and abuts the extracellular fluid. If the end at $x = 0$ is killed, then for $x < 0$ the depolarization is zero, as is the resting potential. This killed end is sometimes referred to as a short-circuit termination.

The initial data describe the depolarization present at the beginning of the experiment for all relevant values of x. Thus we have

$$u(x,0) = s(x), \qquad 0 \le x \le L. \tag{6}$$

In the next section we will introduce a numerical scheme to approximate the solution of general problems of the form

$$\frac{\partial^\alpha u(t,x)}{\partial t^\alpha} = \frac{\partial^2 u(t,x)}{\partial x^2} - u(t,x) + v(t,x), \quad t \in (0,T), \quad x \in (0,L) \tag{7}$$

which satisfies the boundary conditions (4), (5) and the initial condition (6), and where $\dfrac{\partial^\alpha u(t,x)}{\partial t^\alpha}$ represents the Caputo derivative of order α, $0 < \alpha < 1$ of the function u with respect to the variable t, which, for the considered values of α is given by:

$$\frac{\partial^\alpha u(t,x)}{\partial t^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} \frac{\partial u(s,x)}{\partial s} ds, \; t > 0.$$

2 Numerical Method

In order to discretize (7), we will consider a time graded mesh defined through the meshpoints $t_i = \left(\frac{i}{N}\right)^r T$, $i = 0, 1, \ldots, N$, where $r \geq 1$ is the grading exponent, and a uniform space mesh with stepsize $h = \frac{L}{M}$. It should be noticed that when $r = 1$ we obtain a uniform time mesh; if not, the length of each time interval of the partition is variable and denoted by $\tau_i = t_{i+1} - t_i$, $i = 0, ..., N-1$. At each point (t_i, x_j), $i = 1, \ldots, N$, $j = 1, \ldots, M-1$, we are going to use the L1 approximation formula for the fractional time derivative (see for example [4])

$$\frac{\partial^\alpha u(t_i, x_j)}{\partial t^\alpha} \approx \sum_{k=0}^{i-1} b^i_{k+1}(u(t_{k+1}, x_j) - u(t_k, x_j)), \tag{8}$$

where

$$b^i_{k+1} = \frac{1}{\Gamma(2-\alpha)\tau_k}((t_i - t_k)^{1-\alpha} - (t_i - t_{k+1})^{1-\alpha}), \tag{9}$$

and a central finite difference formula for the space derivative:

$$\frac{\partial^2 u(t_i, x_j)}{\partial x^2} \approx \frac{u(t_i, x_{j+1}) - 2u(t_i, x_j) + u(t_i, x_{j-1})}{h^2}. \tag{10}$$

It is known that if $u \in C^{2,4}$ $([0, T] \times [0, L])$, then the order of the approximations (8) and (10) are $2 - \alpha$ and 2, respectively, but it is also known that the class of problems which satisfies such regularity assumptions with respect to the time variable is very restrictive [3]. Moreover, in [7], the authors proved that typical solutions of problems of the form (4)–(7) satisfy:

$$\left|\frac{\partial^k u}{\partial x^k}(t, x)\right| \leq C \quad for \quad k = 0, 1, 2, 3, 4, \tag{11}$$

$$\left|\frac{\partial^l u}{\partial t^l}(t, x)\right| \leq C(1 + t^{\alpha-l}) \quad for \quad l = 0, 1, 2, \tag{12}$$

for all $(t, x) \in (0, T] \times [0, L]$.

In Theorem 1 of [6] it is also shown that if the solution of such problems is less singular than these typical solutions, in the sense that

$$\left|\frac{\partial^l u}{\partial t^l}(t, x)\right| \leq C(1 + t^{\beta-l}) \quad for \quad l = 0, 1, 2, \tag{13}$$

for some $\beta > \alpha$, then the initial condition $s(x)$ is uniquely defined by the right-hand side function $v(t, x)$.

Inserting (8) and (10) in (7), taking into account the initial and boundary conditions and denoting by $u^i_j \approx u(t_i, x_j)$, $v^i_j = v(t_i, x_j)$ we obtain the finite difference scheme

$$u_j^i(h^2b_i^i+2+h^2)-u_{j+1}^i-u_{j-1}^i = = h^2b_i^iu_j^{i-1}-h^2\sum_{k=0}^{i-2}(b_{k+1}^iu_j^{k+1}-u_j^k)+v_j^ih^2,$$

$i=1,\dots,N,\ \ j=1,\dots,M-1$, which in a matrix form writes

$$\begin{pmatrix} h^2b_i^i+2+h^2 & -1 & 0 & \dots & 0 \\ -1 & h^2b_i^i+2+h^2 & -1 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & \dots & -1 & h^2b_i^i+2+h^2 \end{pmatrix}\begin{pmatrix} u_1^i \\ u_2^i \\ \vdots \\ u_{M-1}^i \end{pmatrix} = \tag{14}$$

$$= h^2b_i^i\begin{pmatrix} u_1^{i-1} \\ u_2^{i-1} \\ \vdots \\ u_{M-1}^{i-1} \end{pmatrix} - h^2\begin{pmatrix} \sum_{k=0}^{i-2}b_{k+1}^i(u_1^{k+1}-u_1^k) \\ \sum_{k=0}^{i-2}b_{k+1}^i(u_2^{k+1}-u_2^k) \\ \vdots \\ \sum_{k=0}^{i-2}b_{k+1}^i(u_{M-1}^{k+1}-u_{M-1}^k) \end{pmatrix} + h^2\begin{pmatrix} v_1^i \\ v_2^i \\ \vdots \\ v_{M-1}^i \end{pmatrix} + \begin{pmatrix} u_0^i \\ 0 \\ \vdots \\ u_M^i \end{pmatrix}$$

Before we proceed with the analysis of this numerical scheme, we will prove some auxiliary results.

Lemma 1. *Assume that* $0<\alpha<1$. *The coefficients* b_k^i, $i,k=1,2,\dots N$, *defined in (9), satisfy*

1. $b_i^i>0$, $i=1,2,\dots,N$,
2. $b_i^i>b_{i+1}^{i+1}$, $i=1,2,\dots,N-1$,
3. $b_{k+1}^i>b_k^i$, $k=1,\dots,i$ *and* $i=1,\dots,N$,
4. $b_1^j\geq b_1^l$, $1\leq j\leq l$ *and* $l=1,\dots,N$,
5. $t_j^\alpha b_1^j\geq t_l^\alpha b_1^l$, $1\leq j\leq l$, $l=1,\dots,N$.

Proof. 1. $b_i^i=\dfrac{1}{\Gamma(2-\alpha)\tau_{i-1}}((t_i-t_{i-1})^{1-\alpha}=\dfrac{1}{\Gamma(2-\alpha)\tau_{i-1}}(\tau_{i-1})^{1-\alpha}>0.$

2. $b_i^i=\dfrac{1}{\Gamma(2-\alpha)\tau_{i-1}}(\tau_{i-1})^{1-\alpha}>\dfrac{1}{\Gamma(2-\alpha)\tau_i}(\tau_i)^{1-\alpha}=b_{i+1}^{i+1}.$
3. To prove this, we will use the definition of b_{k+1}^i in (9):

$$b_{k+1}^i=\frac{1}{\Gamma(1-\alpha)\tau_k}\int_{t_k}^{t_{k+1}}(t_i-s)^{-\alpha}ds. \tag{15}$$

So, by the theorem of the mean value for integrals, there must exist a $\kappa_1^k\in(t_k,t_{k+1})$ such that

$$b_{k+1}^i=\frac{1}{\Gamma(1-\alpha)\tau_k}(t_i-\kappa_1^k)^{-\alpha}\tau_k=\frac{1}{\Gamma(1-\alpha)}(t_i-\kappa_1^k)^{-\alpha}.$$

Following the same idea, there must exist a $\kappa_2^k\in(t_{k-1},t_k)$ such that

$$b_k^i = \frac{1}{\Gamma(1-\alpha)\tau_{k-1}}(t_i - \kappa_2^k)^{-\alpha}\tau_{k-1} = \frac{1}{\Gamma(1-\alpha)}(t_i - \kappa_2^k)^{-\alpha}.$$

Since $\kappa_2^k < \kappa_1^k$, we have the following

$$b_{k+1}^i - b_k^i = \frac{1}{\Gamma(1-\alpha)}((t_i - \kappa_1^k)^{-\alpha} - (t_i - \kappa_2^k)^{-\alpha}) > 0. \tag{16}$$

4. We must prove that

$$\begin{aligned}&\left((t_j - t_0)^{1-\alpha} - (t_j - t_1)^{1-\alpha}\right) \geq \left((t_l - t_0)^{1-\alpha} - (t_l - t_1)^{1-\alpha}\right) \Leftrightarrow \\ &\left(t_j^{1-\alpha} - (t_j - t_1)^{1-\alpha}\right) \geq \left(t_l^{1-\alpha} - (t_l - t_1)^{1-\alpha}\right)\end{aligned}$$

Let $f(x) = x^{1-\alpha} - (x - t_1)^{1-\alpha}$. In this case, the inequality above writes $f(t_j) - f(t_l) \geq 0$. Using the mean value theorem, we have $f(t_j) - f(t_l) = (t_j - t_l)$ $f'(\eta)$, $\eta \in]t_j, t_l[$. Note that $t_j - t_l \leq 0$. Since for $x > t_1$,

$$f'(x) = (1-\alpha)x^{-\alpha} - (1-\alpha)(x - t_1)^{-\alpha} = (1-\alpha)(x^{-\alpha} - (x - t_1)^{-\alpha}) < 0$$

we conclude that $f(t_j) - f(t_l) \geq 0$, $1 \leq j \leq l$, $l = 1, ..., N$.

5. Here we must prove that

$$t_j^{\alpha}\left\{(t_j - t_0)^{1-\alpha} - (t_j - t_1)^{1-\alpha}\right\} \geq t_l^{\alpha}\left\{(t_l - t_0)^{1-\alpha} - (t_l - t_1)^{1-\alpha}\right\}$$

Proceeding as in 5, it suffices to show that $g'(x) \leq 0$ $\forall x > t_1$ for $g(x) = x^{\alpha}(x^{1-\alpha} - (x - t_1)^{1-\alpha})$. In fact,

$$\begin{aligned}g'(x) &= \alpha x^{\alpha-1}(x^{1-\alpha} - (x - t_1)^{1-\alpha}) + (1-\alpha)x^{\alpha}(x^{-\alpha} - (x - t_1)^{-\alpha}) \\ &= 1 - \alpha x^{\alpha-1}(x - t_1/x)^{1-\alpha} - (1-\alpha)(1 - t_1/x)^{-\alpha}\end{aligned}$$

Let $t_1 < x$, By the Newton Binomial we get the following:

$$\begin{aligned}\left(1 - \frac{t_1}{x}\right)^{1-\alpha} &= 1 - \frac{t_1}{x}(1-\alpha) + \left(\frac{t_1}{x}\right)^2(1-\alpha)\left(\frac{-\alpha}{2}\right) + O\left(\frac{t_1}{x}\right)^2 \\ \Leftrightarrow \left(1 - \frac{t_1}{x}\right)^{1-\alpha} &= 1 + \frac{t_1}{x}\alpha + \left(\frac{t_1}{x}\right)^2\frac{(-\alpha)(1-\alpha)}{2} + O\left(\frac{t_1}{x}\right)^2\end{aligned}$$

$$\alpha\left(1 - \frac{t_1}{x}\right)^{1-\alpha} = \alpha - \alpha(1-\alpha)\frac{t_1}{x} + \frac{\alpha^2}{2}(1-\alpha)\left(\frac{t_1}{x}\right)^2 + o\left(\frac{t_1}{x}\right)^2 \tag{17}$$

$$(1-\alpha)\left(1 - \frac{t_1}{x}\right)^{-\alpha} = (1-\alpha) - \alpha(1-\alpha)\frac{t_1}{x} + (1-\alpha)\frac{\alpha(1+\alpha)}{2}\left(\frac{t_1}{x}\right)^2 + o\left(\frac{t_1}{x}\right)^2$$

Taking (17) into account, we obtain:

$$\begin{aligned} g'(x) &= 1-\Big(\alpha-\alpha(1-\alpha)\frac{t_1}{x}+\frac{\alpha^2}{2}(1-\alpha)\Big(\frac{t_1}{x}\Big)^2+o\Big(\frac{t_1}{x}\Big)^2 \\ &+ (1-\alpha)-\alpha(1-\alpha)\frac{t_1}{x}+(1-\alpha)\frac{\alpha(1+\alpha)}{2}\Big(\frac{t_1}{x}\Big)^2+o\Big(\frac{t_1}{x}\Big)^2\Big) \\ &= 1-\alpha-1+\alpha-\alpha(1-\alpha)\frac{t_1}{x}+\alpha(1-\alpha)\frac{t_1}{x}-\frac{1-\alpha}{2}\alpha\Big(\frac{t_1}{x}\Big)^2+o\Big(\frac{t_1}{x}\Big)^2 \\ &\to g'(x)<0 \end{aligned}$$

Lemma 2. *[3] Let*

$$S^i=\delta_t^\alpha u^i-\frac{d^\alpha}{dt^\alpha}u(t_i),\ i=1,\ldots,N, \tag{18}$$

where $\delta_t^\alpha u^i$ denotes an approximation of $u(t_i)$ given by the L1 formula (see (8)). Then

$$|S^i|\le t_i^{-\alpha}\max_{k=1,\ldots,i}\phi^k,$$

where

$$\begin{aligned} \phi^1 &= \tau_1^\alpha \sup_{s\in(0,t_1)}\left(s^{1-\alpha}\left|\frac{u(t_1)-u(t_0)}{t_1-t_0}-\frac{d}{ds}u(s)\right|\right), \\ \phi^k &= \tau_k^{2-\alpha}t_k^\alpha \sup_{s\in(t_{k-1},t_k)}\left|\frac{d^2}{ds^2}u(s)\right|,\ k=2,\ldots,i. \end{aligned}$$

Lemma 3. *If a function $u(t)$ satisfies*

$$u^{(\ell)}(t)\le C(1+t^{\beta-\ell})\,for\ \ell=0,1,2, \tag{19}$$

for some $\beta\ge\alpha$, then

$$|S^i|\le t_i^{-\alpha}N^{-\min\{\beta r,2-\alpha\}},\ i=1,\ldots,N.$$

Proof. It suffices to show that

$$\phi^j\le CN^{-min\{\beta r,2-\alpha\}},\, j\ge 1 \tag{20}$$

Let us first analyse ϕ^1. Note that

$$\phi^1\le\tau_1^\alpha\left[\sup_{s\in(0,t_1)}\left(s^{1-\alpha}\left|\frac{u(t_1)-u(t_0)}{t_1-t_0}\right|\right)+\sup_{s\in(0,t_1)}\left(s^{1-\alpha}\left|\frac{d}{ds}u(s)\right|\right)\right].$$

Since

$$s^{1-\alpha}\left|\frac{u(t_1)-u(t_0)}{t_1-t_0}\right| \leq \tau_1^{-\alpha}\Big|\int_{t_0}^{t_1} u'(s)ds\Big| \leq C\tau_1^{-\alpha}\Big|\int_0^{t_1} s^{\beta-1}ds\Big|$$
$$= C\frac{\tau_1^{-\alpha}}{\beta}\tau_1^{\beta} \leq D\tau_1^{\beta-\alpha},$$

and

$$s^{1-\alpha}\left|\frac{d}{ds}u(s)\right| \leq Cs^{1-\alpha}s^{\beta-1} = Cs^{\beta-\alpha} \leq C_1\tau_1^{\beta-\alpha}.$$

we conclude (because $\tau_1 = t_1 \leq CN^{-r}$) that

$$\phi^1 \leq \overline{C}\tau_1^{\alpha}\tau_1^{\beta-\alpha} = \overline{C}\tau_1^{\beta} \leq \overline{C}N^{-\beta r}.$$

Let us now analyse ϕ^j, $j \geq 2$. From (19) we have

$$\sup_{s\in(t_{j-1},t_j)}\left|\frac{d^2}{ds^2}u(s)\right| \leq Ct_j^{\beta-2}.$$

Then

$$\phi^j \leq C\tau_j^{2-\alpha}t_j^{\alpha}t_j^{\beta-2} = C\left(\frac{\tau_j}{t_j}\right)^{2-\alpha}t_j^{\beta},\ j \geq 2$$

Let $\gamma = min\{\beta r, 2-\alpha\}$. Then

$$\phi^j \leq C\left(\frac{\tau_j}{t_j}\right)^{\gamma}t_j^{\beta},\ j \geq 2 \leq \tilde{C}\frac{N^{-\gamma}t_j^{\gamma(1-1/r)}}{t_j}t_j^{\beta} = \tilde{C}N^{-\gamma}t_j^{-\gamma/r+\beta}$$

Since $-\gamma/r+\beta \geq 0$, then

$$\phi^j \leq N^{-\gamma}t_j^{-\gamma/r+\beta} \leq DN^{-\gamma} = DN^{-min\{\beta r,2-\alpha\}},\ j \geq 2.$$

Theorem 1. *The numerical scheme (14) is uniquely solvable.*

Proof. Taking into account Lemma 1, we have $b_i^i > 0$ for each $i = 1, ..., N$, and then the matrix in the left-hand side of (14) is strictly diagonal dominant, and therefore, the system has a unique solution.

In order to prove the stability and convergence of the numerical scheme, let us first note that (2) can be rewritten as

$$L_1u_j^i = L_2u_j^{i-1} + v_j^ih^2 \tag{21}$$

where

$$L_1 u_j^i = u_j^i (h^2 b_i^i + 2 + h^2) - u_{j+1}^i - u_{j-1}^i \tag{22}$$

$$L_2 u_j^{i-1} = h^2 b_i^i u_j^{i-1} - h^2 \sum_{k=0}^{i-2} b_{k+1}^i (u_j^{k+1} - u_j^k). \tag{23}$$

Let's consider the initial condition has the error as ϵ_j^0, that is, let's replace $s(x_j)$ with $\widetilde{s}(x_j) = s(x_j) + \epsilon_j^0$, $j = 1, ..., M-1$ and let u_j^i and $\widetilde{u}_j^i$ be the solutions of the problem (4)–(7) with initial condition s and $\widetilde{s}$, respectively. Defining the error as $\epsilon_j^i = u_j^i - \widetilde{u}_j^i$, we have the following

$$L_1 \epsilon_j^i = L_2 \epsilon_j^{i-1}. \tag{24}$$

Theorem 2. *The numerical scheme (2) is unconditionally stable.*

Proof. We will prove that $\left\| E^{k+1} \right\| \leq ||E^0||$, $k = 0, 1, 2, ..., N-1$, where $E^k = [\epsilon_1^k, \epsilon_2^k, ..., \epsilon_{M-1}^k]^T$ by using mathematical induction. For $k = 0$ and assuming $|\epsilon_l^1| = \max_{j=1,...,M} |\epsilon_j^1|$, we have

$$\begin{aligned} h^2 b_1^1 ||E^1|| &= h^2 b_1^1 |\epsilon_l^1| \leq h^2 b_1^1 |\epsilon_l^1| + h^2 |\epsilon_l^1| = h^2 b_1^1 |\epsilon_l^1| + h^2 |\epsilon_l^1| + 2|\epsilon_l^1| - 2|\epsilon_l^1| \\ &\leq h^2 b_1^1 |\epsilon_l^1| + h^2 |\epsilon_l^1| + 2|\epsilon_l^1| - |\epsilon_{l+1}^1| - |\epsilon_{l-1}^1| \\ &= |\epsilon_l^1|(h^2 b_1^1 + 2 + h^2) - |\epsilon_{l+1}^1| - |\epsilon_{l-1}^1| \leq |\epsilon_l^1 (h^2 b_1^1 + 2 + h^2) - \epsilon_{l+1}^1 - \epsilon_{l-1}^1| \\ &= |L_1 \epsilon_l^1| = |L_2 \epsilon_l^0| = |h^2 b_1^1 \epsilon_l^0| = h^2 b_1^1 |\epsilon_l^0| \leq h^2 b_1^1 ||E^0||, \end{aligned}$$

and therefore $||E^1|| \leq ||E^0||$. Now suppose that $||E^k|| \leq ||E^0||$, with $k = 0, 1, 2, ..., i$ and $i = 1, ..., N-1$. Let also $|\epsilon_l^{k+1}| = \max_{j=1,...,M-1} |\epsilon_j^{k+1}|$. Using again Lemma 1 and the triangular inequality we are going to prove that $||E^{k+1}|| \leq ||E^0||$.

$$\begin{aligned} h^2 b_{k+1}^{k+1} ||E^{k+1}|| &= h^2 b_{k+1}^{k+1} |\epsilon_l^{k+1}| \leq h^2 b_{k+1}^{k+1} |\epsilon_l^{k+1}| + h^2 |\epsilon_l^{k+1}| + 2|\epsilon_l^{k+1}| - 2|\epsilon_l^{k+1}| \\ &\leq |\epsilon_l^{k+1}(h^2 b_{k+1}^{k+1} + 2 + h^2) - \epsilon_{l+1}^{k+1} - \epsilon_{l-1}^{k+1}| = |L_1 \epsilon_l^{k+1}| = |L_2 \epsilon_l^k| \\ &= |h^2 b_{k+1}^{k+1} \epsilon_l^k - h^2 \sum_{s=0}^{k-1} b_{s+1}^{k+1} (\epsilon_l^{s+1} - \epsilon_l^s)| \\ &= |h^2 b_1^{k+1} \epsilon_l^0 + h^2 \sum_{s=0}^{k} (b_{s+1}^{k+1} - b_s^{k+1}) \epsilon_l^{k+1}| \\ &\leq h^2 b_1^{k+1} |\epsilon_l^0| + h^2 \sum_{s=1}^{k} (b_{s+1}^{k+1} - b_s^{k+1}) |\epsilon_l^k| \\ &\leq h^2 b_1^{k+1} ||E^0|| + h^2 \sum_{s=1}^{k} (b_{s+1}^{k+1} - b_s^{k+1}) ||E^0|| = h^2 b_{k+1}^{k+1} ||E^0||, \end{aligned}$$

and then $||E^{K+1}|| \leq ||E^0||$.

For the convergence analysis of the numerical scheme we define the errors at the mesh points:

$$\eta_j^i = u(t_i, x_j) - u_j^i \tag{25}$$

for $i = 1, ..., N$, $j = 1, ..., M - 1$ and we consider the error vector at the time-level i:

$$H^i = (\eta_1^i, \eta_2^i, ..., \eta_{M-1}^i)^T, \ i = 1, ..., N. \tag{26}$$

Taking into account the approximations for the time and space derivatives, the differential equation in (7) may be written at the point $(t, x) = (t_i, x_j)$ as:

$$\sum_{k=0}^{i-1} b_{k+1}^i (u(t_{k+1}, x_j) - u(t_k, x_j)) = \frac{u(t_i, x_{j+1}) - 2u(t_i, x_j) + u(t_i, x_{j-1})}{h^2} - u(t_i, x_j) \\ + v(t_i, x_j) + R^{ij} \tag{27}$$

where $i = 1, ..., N$, $j = 1, ..., M - 1$ and R^{ij} comprises of the errors committed in the approximations of the time Caputo derivative and the space second order derivative We will assume that the solution of (4)–(7) satisfies conditions (11) and (13), for $\beta \geq \alpha$, and then for $i = 1, ..., N, \quad j = 1, ..., M - 1$,

$$||R^{ij}||_\infty = \max_{j=1,...,M-1} |R^{ij}| \leq C(t_i^{-\alpha} N^{-min\{\beta r, 2-\alpha\}} + h^2) = R^i \tag{28}$$

We easily see that the errors η_j^i satisfy:

$$\begin{cases} L_1 \eta_j^{i+1} = L_2 \eta_j^i + h^2 R^{ij}, \quad i = 1, 2, ..., N - 1, \quad j = 1, 2, ..., M - 1 \\ \eta_j^0 = 0, \quad j = 1, 2, ..., M - 1 \end{cases} . \tag{29}$$

with L_1 and L_2 defined in (22) and (23), respectively.

Lemma 4. *Under the conditions (11) and (13), there must exist a positive constant C_1 such that:*

$$||H^l|| \leq \frac{C_1(t_l^{-\alpha} N^{-\gamma} + h^2)}{b_l^l - \sum_{k=0}^{l-2}(b_{k+2}^l - b_{k+1}^l)}, \tag{30}$$

for $l = 1, ..., N$, $\gamma = min\{\beta r, 2 - \alpha\}$ and H^l defined in (26).

Proof. We use mathematical induction to prove the Lemma 4. For $l = 1$, let p to be a natural number such that $||H^1|| = \max_{j=1,...,M} |\eta_j^1| = |\eta_p^1|$. Then,

$$\begin{aligned}
&b_1^1 h^2 ||H^1|| = b_1^1 h^2 |\eta_p^1| \leq b_1^1 h^2 |\eta_p^1| + h^2 |\eta_p^1| + 2|\eta_p^1| - |\eta_p^1| - |\eta_p^1| \\
&\leq (h^2 b_1^1 + h^2 + 2)|\eta_p^1| - |\eta_{p+1}^1| - |\eta_{p-1}^1| \\
&\leq |(h^2 b_1^1 + h^2 + 2)\eta_p^1 - \eta_{p+1}^1 - \eta_{p-1}^1| = \|L_1 \eta_p^1| = |L_2 \eta_p^0 + h^2 R^{1p}| \leq h^2 ||R^{1p}||_\infty \\
&\leq C_1 (t_l^{-\alpha} N^{-\gamma} + h^2)
\end{aligned}$$

and then (30) is proved for $l = 1$. Let us assume that

$$||H^m|| \leq \frac{C_1(t_m^{-\alpha} N^{-\gamma} + h^2)}{b_m^m - \sum_{k=0}^{m-2}(b_{k+2}^m - b_{k+1}^m)},$$

holds for $m = 1, ..., l-1$, and show that it remains valid for $m = l$. Let $q \in \mathbb{N}$ be such that $||H^l||_\infty = |\eta_q^l|$.

$$\begin{aligned}
h^2 b_l^l ||H^l|| &= h^2 b_l^l |\eta_q^l| \leq h^2 b_l^l |\eta_q^l| + h^2 |\eta_q^l| = h^2 b_l^l |\eta_q^l| + h^2 |\eta_q^l| + 2|\eta_q^l| - 2|\eta_q^l| \\
&\leq |(h^2 b_l^l + h^2 + 2)\eta_q^l + \eta_q^l - \eta_{q+1}^l - \eta_{q-1}^l| = |L_1 \eta_q^l| \\
&= |L_2 \eta_q^{l-1} + h^2 R^{lq}| \leq |L_2 \eta_q^{l-1}| + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\
&= |h^2 b_l^l |\eta_q^{l-1} - h^2 \sum_{k=0}^{l-2} b_{k+1}^l (\eta_q^{k+1} - \eta_q^k)| + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\
&= h^2 \sum_{k=1}^{l-1} (b_{k+1}^l - b_k^l)|\eta_q^k| + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\
&= h^2 \sum_{j=0}^{l-2} (b_{j+2}^l - b_{j+1}^l)|\eta_q^{j+1}| + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\
&= h^2 \sum_{j=0}^{l-2} (b_{j+2}^l - b_{j+1}^l)||H^{j+1}|| + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\
&\leq h^2 \sum_{j=0}^{l-2} (b_{j+2}^l - b_{j+1}^l) \frac{C_1(t_{j+1}^{-\alpha} N^{-\gamma} + h^2)}{b_{j+1}^{j+1} - \sum_{k=0}^{j-2}(b_{k+2}^{j+1} - b_{k+1}^{j+1})} + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2)
\end{aligned}$$

where in the last step, we used the induction hypothesis.

Defining

$$\mathbb{A}_l = b_l^l - \sum_{j=0}^{l-2}(b_{j+2}^l - b_{j+1}^l), \tag{31}$$

we easily verify that $\mathbb{A}_l = b_1^l$ and then, using the two last properties of Lemma 1,

$$\begin{aligned} b_l^l h^2 ||H^l||_\infty &\leq h^2 \sum_{j=0}^{l-2} (b_{j+2}^l - b_{j+1}^l) \frac{C_1(t_{j+1}^{-\alpha} N^\gamma + h^2)}{b_1^{j+1}} + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\ &\leq h^2 \sum_{j=0}^{l-2} (b_{j+2}^l - b_{j+1}^l) \frac{C_1(t_{j+1}^{-\alpha} N^\gamma + h^2)}{b_1^l} + Ch^2(t_l^{-\alpha} N^{-\gamma} + h^2) \\ &\leq \frac{\bar{C}_1 h^2 b_l^l (t_1^{-\alpha} N^{-\gamma} + h^2)}{\mathbb{A}_l} \end{aligned} \tag{32}$$

the result is proved.

Next we prove the main result concerning the convergence analysis of the numerical scheme.

Theorem 3. *Assume that the solution of (4)–(7) satisfies (11) and (13). Then*

$$||H^l||_\infty \leq C(N^{-\gamma} + h^2), \quad l = 1, ..., N, \tag{33}$$

where $\gamma = min\{\beta r, 2 - \alpha\}$.

Proof. First note that

$$\mathbb{A}_l = b_1^l = \frac{\tau_0^{-\alpha}}{\Gamma(2-\alpha)} \left((l^r)^{1-\alpha} - (l^r - 1)^{1-\alpha} \right) \tag{34}$$

Since

$$\lim_{l\to\infty} \frac{(l^r)^{-\alpha}}{(l^r)^{1-\alpha} - (l^r - 1)^{1-\alpha}} = \lim_{\delta\to\infty} \frac{\delta^{-\alpha}}{\delta^{1-\alpha} - (\delta - 1)^{1-\alpha}} = \frac{1}{1-\alpha}. \tag{35}$$

then, there must exist a positive constant C_2 such that for large l, (30) becomes

$$\begin{aligned} ||H^l|| \leq \frac{C_1 C_2 (t_l^{-\alpha} N^{-\gamma} + h^2)}{(l^r)^{-\alpha} \tau_0^{-\alpha}} &= \frac{C_1 C_2 (t_l^{-\alpha} N^{-\gamma} + h^2)}{(l^r t_1)^{-\alpha}} = \frac{C_1 C_2 (t_l^{-\alpha} N^{-\gamma} + h^2)}{t_l^{-\alpha} \tau_0^{-\alpha}} \\ = C_1 C_2 N^{-\gamma} + t_l^\alpha h^2 &\leq C(N^{-\gamma} + h^2) = C(N^{-min(\beta r, 2-\alpha)} + h^2). \end{aligned} \tag{36}$$

3 Numerical Results and Discussion

In this case, we do not know the exact solution to compare it with the numerical results. For this reason we introduce the following estimate of the convergence order with respect to time:

$$k = log_2 \left(\frac{|U^N - U^{2N}|}{|U^{2N} - U^{4N}|} \right) \tag{37}$$

where U^N stands for the solution of the Eq. (7) obtained with N steps in time.

Table 1 Solution of the killed end problem with $M = 100$ at the point $(\Upsilon, \chi) = (1, 0.5)$, using different uniform meshes in time. k is the estimate of the convergence order with respect to t (see (37))

α	0.2		0.4		0.6		0.8	
N	U_M^N	k	U_M^N	k	U_M^N	k	U_M^N	k
100	345.362		350.825		357.585		365.654	
200	345.375		350.848		357.613		365.68	
400	345.382	1.0051	350.86	1.009	357.627	1.019	365.692	1.04
800	345.386	1.0027	350.866	1.0052	357.634	1.0127	365.698	1.03

Table 2 Solution of the problem killed end problem with $M = 100$ at the point $(\Upsilon, \chi) = (1, 0.5)$, using a non uniform mesh in time with $r = \frac{2-\alpha}{\alpha}$. k is the estimate of the convergence order with respect to t (see (37))

α	0.2		0.4		0.6		0.8	
N	U_M^N	k	U_M^N	k	U_M^N	k	U_M^N	k
100	345.44		350.87		357.634		365.681	
200	345.523		350.871		357.638		365.694	
400	345.533	1.82	350.872	1.62	357.64	1.43	365.699	1.19
800	345.513	1.81	350.872	1.61	320.766	1.42	365.702	1.18

Table 3 Solution of the killed end problem with $N = 100$ at the point $(\Upsilon, \chi) = (0.5, 1)$, using a uniform mesh in space. p is the estimate of the convergence order with respect to x (see (38))

α	0.2		0.4		0.6		0.8	
M	U_M^N	p	U_M^N	p	U_M^N	p	U_M^N	p
100	345.362		350.825		357.585		365.654	
200	345.363		350.826		357.586		365.655	
400	345.364	2.0	350.827	2.0	357.587	2.0	365.655	2.0
800	345.364	2.0	350.827	2.0	357.587	2.0	365.693	2.0

We also present an estimate of the convergence order with respect to space stepsize, h, similar to (37):

$$p = log_2\left(\frac{|U_M - U_{2M}|}{|U_{2M} - U_{4M}|}\right) \tag{38}$$

where U_M stands for the numerical solution obtained with M steps in space.

We consider the initial condition $s(x) = -70\chi(\chi - 1)mV$. The following problem is the one that we are going to solve numerically using the the method described in Sect. 2.

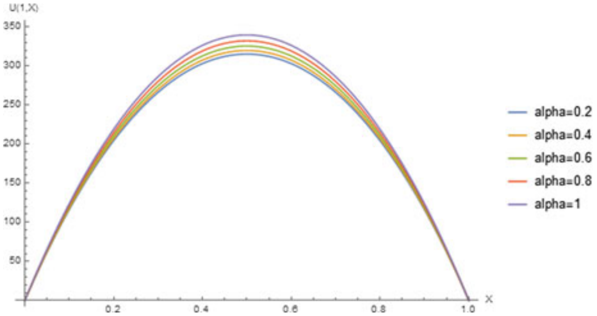


Fig. 2 The membrane potential, $U(1, \chi)$, at the fixed point $\Upsilon = 1$ for different values of $\alpha \in]0, 1[$ using a time non-uniform and space uniform mesh (N = 400 and M = 100)

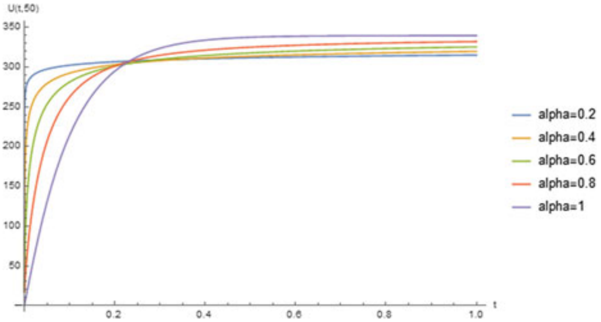


Fig. 3 The membrane potential, $U(\Upsilon, M/2)$, at the fixed space point $\chi = M/2$ for different values of $\alpha \in]0, 1[$ using a space uniform and a time non-uniform mesh (N = 400 and M = 100)

$$\begin{array}{c} \frac{\partial^\alpha U(\Upsilon,\chi)}{\partial \Upsilon^\alpha} = \frac{\partial^2 U(\Upsilon,\chi)}{\partial \chi^2} - U(\Upsilon, \chi) + 300, \quad \Upsilon \in [0, 1], \quad \chi \in [0, 1], \\ U(0, \chi) = -70\chi(\chi - 1), \quad \chi \in]0, 1[, \\ U(\Upsilon, 0) = 0, \quad \Upsilon \in [0, 1], \\ U(\Upsilon, 1) = 0, \quad \Upsilon \in [0, 1]. \end{array} \tag{39}$$

In Table 1 we are going to present the solution of the problem (39) in a single point, $U(t_i, x_j)$, where $i = N$ and $j = M/2$ because this is where the biggest errors are expected (taking into account the usual behaviour of the solution in diffusion problems with this kind of boundary conditions).

The results in Tables 1, 2 and 3 confirm the theoretical results obtained in Sect. 2. When we calculate the solution using a uniform mesh we obtain a experimental order of $k \approx 1$ (see Table 1), while with the non uniform mesh we get the expected experimental order $k \approx 2 - \alpha$ with $\alpha \in]0, 1[$ (see Table 2), when using $r = \frac{2-\alpha}{\alpha}$. The numerical results confirm that the method is convergent and the experimental order is in agreement with the theoretical predictions in Sect. 2. In particular, when a graded mesh is used, with the proper value of the grading coefficient, the optimal convergence order $k = 2 - \alpha$ is recovered, in spite of the singularity of the problem.

In Table 3 we calculate the space convergence order to show that it is also in agreement with the theoretical results in Sect. 2.

Besides (39), we decided to consider an analogous problem with the first order temporal derivative instead of the fractional temporal one, and check the difference between these two approaches.

To obtain the numerical results in the case of the integer order problem, we used the implicit Euler method for the time derivative and the central finite difference method for the second order space derivative.

Some graphics of the numerical results obtained by this scheme are plotted in Figs. 2 and 3.

In Fig. 2 we display the solution profile at a fixed time moment ($\Upsilon = 1$). As it could be expected the maximal values of the solution are attained close to the midpoint ($\chi = 0.5$) and the solution with $\alpha = 1$ (first order derivative) has higher values than the other ones.

Figure 3 illustrates how the solution changes with time, with $\chi = 0.5$, and how its behavior depends on α. For values of Υ close to 0, the solution with $\alpha = 1$ (integer order) has smaller values than the solutions with other values of α. We can also observe that this solution has a regular behavior near the origin, while as α decreases the slope of the graphic at the origin becomes steeper and steeper. These results are in agreement with what could be expected, knowing the asymptotic behavior of the solutions of fractional order equations.

4 Conclusion

We have introduced a numerical scheme for solving the time fractional cable equation, based on the use of the $L1$-method to approximate the fractional time derivative and a second-order finite difference scheme to approximate the space derivative. We have analysed the stability and convergence of the method, and proved that using a graded mesh, with a proper value of the grading coefficient, it is possible to recover the optimal convergence order of the $L1$-method, in spite of the singularity of the solution at the origin. Finally, we have presented some numerical examples that confirm the theoretical predictions.

References

1. Henry, B.I., Langlands, T.A.M., Wearne, S.L.: Fractional cable models for spiny neuronal dendrites. Phys. Rev. Lett. **100**(12), 128103 (2008)
2. Koch, C.: Biophysics of Computation: Information Processing in Single Neuron, vol. 298–300, pp. 43–44. Oxford University Press, New York (2004)
3. Kopteva, N.: Error analysis of the L1 method on graded and uniform meshes for a fractional derivative problem in two and three dimensions. Math. Comput. **88**, 1–20 (2017)

4. Li, C., Zeng, F.: Numerical Methods for Fractional Calculus, pp. 43–46. Chapman Hall Crcs, Boca Raton (2015)
5. Norman, R.S.: Cable theory for finite length dendritic cylinders with initial and boundary conditions. Biophys. J. **12**(1), 25–45 (1972)
6. Stynes, M.: Too much regularity may force too much uniqueness. Fract. Calc. Appl. Anal. **19**, 1554–1562 (2016)
7. Stynes, M., O'Riordan, E., Gracia, J.L.: Error analysis of a finite difference method on a graded meshes for a time-fraction diffusion equation. SIAM J. Numer. Anal. **55**(2), 1057–1079 (2017)
8. Teka, W., Marinov, T.M.: Neuronal spike timing adaptation described with a fractional leaky integrate-and-fire model. PLoS Comput. Biol. **10**(3), 1003526 (2014)
9. Tuckwell, H.C.: Introduction to Theoretical Neurobiology: Linear Cable Theory and Denditric Structure. Cambridge Studies in Mathematical Biology, vol. 1. Cambridge University Press, Cambridge (1988)
10. Vitali, S., Castellani, G., Mainardi, F.: Time fractional cable equation and applications in neurophysiology. Chaos Solitons Fractals **102**, 467–472 (2017)

Modelling Antibody-Antigen Interactions

Qi Wang and Yupeng Liu

Abstract In this paper we construct and analyse mathematical models for antibody-antigen reactions, which are important for understanding bioaffinity devices. We consider three types of immunoassays: the direct assay, the competitive assay (which are analysed with and without diffusion effects) and the sandwich assay.

1 Introduction

We give examples and analyse problems where modelling of transport phenomena only affect the transient behaviour of the system and has no effect on the final steady states of the species involved. It is often the case that the equilibrium values are the only piece of information required for the solution of a practical problem (although, sometimes, time to reach equilibrium is the real issue) and in such situations it is important to identify the conditions under which a complex partial differential equations model can be replaced with a simpler one. Such problems as these are related to immunosensors, a class of bioaffinity devices, and involve mathematical models of antibody-antigen interactions.

2 The Direct Assay

This section studies the kinetics of the binding reaction between an antigen and an antibody, with and without modelling of transport effects. This simple reaction is rarely used on its own for diagnostic purposes but lies at the heart of every immunosensing device and so we must study it first.

Q. Wang (✉)
School of Hospitality Management and Tourism, Technological University Dublin, Dublin, Ireland
e-mail: qi.wang@TUDublin.ie

Y. Liu
School of Computing, Technological University Dublin, Dublin, Ireland
e-mail: yupeng.liu@TUDublin.ie

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_45

2.1 Simplified Model for the Direct Assay

We start our study of the direct antibody-antigen interactions by ignoring transport of species and concentrating on the kinetics of the reaction. This will result in a simple system of ordinary differential equations model and our aim here is to provide a formula for the equilibrium values of all reactants and products as well as their dependence on initial conditions.

The antibody-antigen interaction can be expressed by the following reaction equation symbolically,

$$A + B \underset{k_-}{\overset{k}{\rightleftarrows}} C, \tag{1}$$

where A represents antigen, B represents antibody, and C represents the product of antigen and antibody. Reaction (1) has a forward (association) reaction rate of k and a backward (dissociation) reaction rate of k_-, where the forward reaction rate is very large (around 1000 times bigger than the reaction rate constant k_1 in the Michaelis-Menten kinetics) while the backward reaction is very slow and is therefore often neglected. This fact reflects the high affinity between antigen and its corresponding antibody. We denote the concentration of the chemical species in reaction (1) by their corresponding lower case letters, namely

$$a = [A], \quad b = [B], \quad c = [C].$$

The dynamics of the system is described by the following non-dimensional system of ordinary differential equations

$$\begin{cases} \frac{da}{dt} = -ab + \mu c & \text{(a)} \\ \frac{db}{dt} = -ab + \mu c & \text{(b)} \\ \frac{dc}{dt} = ab - \mu c, & \text{(c)} \end{cases} \tag{2}$$

with non-dimensional initial conditions $a(0) = \psi$, $b(0) = 1$, $c(0) = 0$, and conservation laws

$$\begin{cases} a + c = \psi & \text{(a)} \\ b + c = 1, & \text{(b)} \end{cases} \tag{3}$$

where

$$\mu = \frac{k_-}{kb_0}, \quad \psi = \frac{a_0}{b_0}. \tag{4}$$

Note that $\mu \ll 1$, since the backward reaction is assumed to be much slower than the forward reaction; also, as soon as the experiment is set up, b_0 is fixed, due to the immobilisation of the antibody.

Based on the non-dimensional conservation laws in (3), we can reduce system (2) down to a single equation in terms of c, then the equilibrium value for c is below, where we must select the root which satisfies the condition $c < 1$.

$$c = \frac{1}{2}\left(1 + \psi + \mu - \sqrt{(1 + \psi + \mu)^2 - 4\psi}\right). \tag{5}$$

This solution is equivalent to that obtained in [1], where a spatially extended model is considered. Note also that if $\mu \ll 1$, the leading order approximation for c is given by

$$c = \frac{1}{2}\left(\psi + 1 - |\psi - 1|\right),$$

which gives different results depending on whether $\psi > 1$ or $\psi < 1$.

In what follows, we derive approximate formulas for the equilibrium values of a, b and c using regular perturbation expansions. Such approximations will allow for a more clear interpretation of these results within the experimental framework. We assume the parameter μ is small and write

$$c = c_0 + \mu c_1 + \mu^2 c_2 + \cdots \tag{6}$$

Collecting coefficients of like powers of μ, at $O(1)$ and $O(\mu)$, we obtain

$$c = 1 + \frac{\mu}{1 - \psi} + \cdots \tag{7}$$

or

$$c = \psi + \frac{\mu\psi}{\psi - 1} + \cdots \tag{8}$$

Now since that the solution of c is less than 1 ($c < a_0$ for the dimensional variables), we have to consider these solutions with regard to the following three cases:

- When $\psi > 1$, we choose the solution
$$c = 1 + \frac{\mu}{1 - \psi} + \cdots;$$
- When $\psi < 1$, we choose the solution
$$c = \psi + \frac{\mu\psi}{\psi - 1} + \cdots;$$
- When $\psi = 1$, we cannot choose either of the two solutions obtained in Eqs. (7) and (8), since the two solutions do not allow $\psi = 1$ (we cannot have a zero denominator). Thus, to obtain the solution in this case, we start the asymptotic analysis again with $\psi = 1$ substituted into the equation in terms of c, which yields
$$c^2 - (2 + \mu)c + 1 = 0. \tag{9}$$

It is now more appropriate to use the expansion

$$c = c_0 + \sqrt{\mu} c_1 + \mu c_2 + \mu\sqrt{\mu} c_3 + \cdots \tag{10}$$

since we can clearly see that there is an $\sqrt{\mu}$ term contained in Eq. (5), and thus we obtain

$$(c_0 + \sqrt{\mu} c_1 + \mu c_2 + \mu\sqrt{\mu} c_3 + \cdots)^2 - (2+\mu)(c_0 + \sqrt{\mu} c_1 + \mu c_2 + \mu\sqrt{\mu} c_3 + \cdots) + 1 = 0.$$

Again, by collecting terms in powers of $\sqrt{\mu}$, at $O(1)$ and $O(\sqrt{\mu})$, we obtain

$$c = 1 - \sqrt{\mu} + \cdots$$

or

$$c = 1 + \sqrt{\mu} + \cdots$$

where $c = 1 + \sqrt{\mu} + \cdots$ cannot be a solution, since $c < 1$.

We now present a summary of the equilibrium values for the antigen, antibody and product in all three cases discussed above.

Case 1: When $a_0 > b_0$ (i.e., $\psi > 1$), the equilibrium solutions are

$$\begin{cases} a = \psi - 1 + \frac{\mu}{\psi - 1} + \cdots & \text{(a)} \\ b = \frac{\mu}{\psi - 1} + \cdots & \text{(b)} \\ c = 1 - \frac{\mu}{\psi - 1} + \cdots & \text{(c)} \end{cases} \tag{11}$$

Case 2: When $a_0 < b_0$ (i.e., $\psi < 1$), the equilibrium solutions are

$$\begin{cases} a = \frac{\mu\psi}{1-\psi} + \cdots & \text{(a)} \\ b = 1 - \psi + \frac{\mu\psi}{1-\psi} + \cdots & \text{(b)} \\ c = \psi - \frac{\mu\psi}{1-\psi} + \cdots & \text{(c)} \end{cases} \tag{12}$$

Case 3: When $a_0 = b_0$ (i.e., $\psi = 1$), the equilibrium solutions are

$$\begin{cases} a = \sqrt{\mu} + \cdots & \text{(a)} \\ b = \sqrt{\mu} + \cdots & \text{(b)} \\ c = 1 - \sqrt{\mu} + \cdots & \text{(c)} \end{cases} \tag{13}$$

In particular, the equilibrium value of the product is

$$c = \begin{cases} 1 - \frac{\mu}{\psi - 1} + \cdots, & \text{if } \psi < 1 \\ \psi - \frac{\mu\psi}{1-\psi} + \cdots, & \text{if } \psi > 1 \\ 1 - \sqrt{\mu} + \cdots, & \text{if } \psi = 1. \end{cases} \tag{14}$$

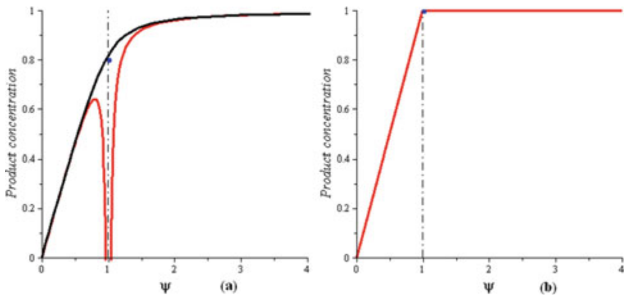


Fig. 1 Product concentration as a function of the initial (non-dimensional) antigen concentration ψ. Black curve correspond to the exact solution of c given by Eq. (5), red curves and the blue dot correspond to the approximate solution of c given by Eq. (14). Typical values for constants used in this simulation are: $b_0 = 2$, $k = 100$, $k_- = 8$ in (a) and $k_- = 0$ in (b)

We note that the asymptotic expansions derived above are not uniformly valid as they fail within an $O(\mu)$ region about $\psi = 1$. (It is easy to see that within this region, the term $\mu/(\psi - 1)$ becomes $O(1)$). Since b_0 is kept constant, we can view c in Eq. (14) as a function of the initial (non-dimensional) antigen concentration $\psi = a_0/b_0$ and this dependence is plotted in Fig. 1, together with the exact solution for c given by Eq. (5). The region of non-uniformity for the asymptotic solution is clearly visible in the figure. However, real immunoassay devices generally work under the condition $a_0 > b_0$ ($\psi > 1$) and in this region we have a uniform approximation. The calibration curve would then consist of the increasing right-hand branch of the red graph in Fig. 1(a). We note that, if $|\psi - 1| > O(\mu)$, then use of the approximation expression (14) might offer better insight into the behaviour of the solution for $\mu \ll 1$, especially for chemistry researchers.

As was expected, the steady states of system (2) depend on whether $a_0 > b_0$ or $a_0 < b_0$ (antigen or antibody predominates). If, for example, the concentration of antigen is greater than that of antibody (Case 1), we see from (11) that, reverting back to dimensional variables, $b \approx 0$, $c \approx b_0$ and $a \approx a_0 - b_0$, which is intuitively clear. (In other words, the antibody is almost depleted and the concentration of product approaches that of the original antibody concentration.) Note also that, if we ignore the backward reaction and let $\mu = 0$, the steady states in this case become: $b^* = 0$, $c^* = b_0$ and $a^* = a_0 - b_0$. Similar interpretations are also easily obtained for the solutions in Cases 2 and 3.

2.2 *Diffusion Model for the Direct Assay*

This subsection covers a spatially extended model of direct antibody-antigen interactions, where the two species are contained within a small cell (which we represent mathematically as a one-dimensional spatial domain). More specifically, we consider the case when the antibody is immobilised on a surface while the antigen is free to diffuse before the interaction between the two species. The resulting model is closely related to the work in [1] and [2], where it was presented as a simplified description (ignoring competitive effects) of a Fluorescence Capillary-Fill Device, a type of pregnancy test studied in [3]. We mention some of the mathematical results obtained in [1] and [2], but the emphasis of this section is on obtaining an exact formula for the equilibrium states of reactants and products and comparing these results to those of the simplified model in Sect. 3.1.

We obtain the non-dimensional system as shown below,

$$\begin{cases} \frac{\partial a(x,t)}{\partial t} = \frac{\partial^2 a(x,t)}{\partial x^2} & \text{(a)} \\ a(x,0) = \psi_1 & \text{(b)} \\ \frac{\partial a(0,t)}{\partial x} = 0 & \text{(c)} \\ \frac{\partial a(1,t)}{\partial x} = \gamma(\mu_1 c(t) - a(1,t)(1-c(t))) & \text{(d)} \\ c(t) + \int_0^1 a(x,t)dx = \psi_1, & \text{(e)} \end{cases} \tag{15}$$

where $x \in (0, 1)$ and we define

$$\psi_1 = \frac{a_0 d}{b_0}, \quad \gamma = \frac{dkb_0}{D}, \quad \mu_1 = \frac{k_- d}{kb_0}. \tag{16}$$

Next, we are going to analyse system (15) as $t \to \infty$. At equilibrium, we obtain the solution of the product concentration as

$$c^* = \frac{1}{2}\left(1 + \mu_1 + \psi_1 - \sqrt{(1 + \mu_1 + \psi_1)^2 - 4\psi_1}\right). \tag{17}$$

Note that the steady-state given above for the diffusion system is identical to the equilibrium value obtained in Eq. (5) for the spatially-independent case, if we allow for the slight differences in the definitions of μ_1, ψ_1 (see Eq. (16)) and μ, ψ (see Eq. (4)).

In what follows, we obtain an equivalent formulation of the diffusion system (15) in the form of a nonlinear Volterra integro-differential equation. We follow the approach suggested in [1] and use Laplace transforms, inverse Laplace transform and convolution theorem for Laplace transform we yields the following Volterra integro-differential equation, namely

$$\frac{dc(t)}{dt} = \gamma\psi_1 - \gamma(\mu_1 + \psi_1)c(t) - \gamma(1 - c(t)) \int_0^t f(t-s)\frac{dc(s)}{ds}ds, \tag{18}$$

where the kernel $f(t)$ can be calculated as

$$\tilde{f}(x,t) = L_t^{-1}\left[\frac{1}{\sqrt{s}} \frac{1 + e^{-2\sqrt{s}x}}{1 - e^{-2\sqrt{s}x}}\right]. \tag{19}$$

Using the geometric series formula

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{n=0}^{\infty} x^n, \text{ for } |x| < 1,$$

we can write Eq. (19) as

$$\tilde{f}(x,t) = L_t^{-1}\left[\frac{1}{\sqrt{\pi}}\sum_{n=0}^{\infty}\left(\sqrt{\frac{\pi}{s}}e^{-2(n+1)\sqrt{s}x} + \sqrt{\frac{\pi}{s}}e^{-2n\sqrt{s}x}\right)\right]. \tag{20}$$

From the theory of Laplace transforms (refer to, for example [4]), we know that

$$L\left[x^{-\frac{1}{2}}e^{-\frac{a}{4x}}\right] = \sqrt{\frac{\pi}{s}}e^{-\sqrt{as}},$$

and hence, we can write Eq. (20) as

$$\tilde{f}(x,t) = \frac{1}{\sqrt{\pi t}}\sum_{n=0}^{\infty}\left(e^{-\frac{(n+1)^2x^2}{t}} + e^{-\frac{n^2x^2}{t}}\right) = \frac{1}{\sqrt{\pi t}}\left(1 + 2\sum_{n=1}^{\infty}e^{-\frac{n^2x^2}{t}}\right),$$

which gives the kernel

$$f(t) = \tilde{f}(1,t) = \frac{1}{\sqrt{\pi t}}\left(1 + 2\sum_{n=1}^{\infty}e^{-\frac{n^2}{t}}\right). \tag{21}$$

We have obtained the integro-differential Eq. (18) as an equivalent formulation for system (15). As illustrated in [1, 2] and [5], this Volterra integro-differential equation is more amenable to both analytical and numerical studies.

In what follows, we find an approximation for $c(t)$, the product concentration, using a regular perturbation method. Consider an analytic expansion for $c(t)$ of the form

$$c(t) = c^0(t) + \varepsilon_2 c^1(t) + \cdots \tag{22}$$

where

$$\varepsilon_2 = \frac{1}{\psi_1} = \frac{b_0}{a_0 d}$$

as suggested in [1], which assumed that γ is order ε, with $\gamma = \gamma\psi_1\varepsilon$, $\gamma\psi_1$ and $\gamma\mu_1$ are of order 1. The non-dimensional parameter ε_2 can be considered small as the antibody sites are usually limited, and is more appropriate for the subsequent perturbation analysis than the parameter μ used previously. Substituting the expansion (22) into (18), at $O(1)$ and $O(\varepsilon_2)$, we obtain

$$\frac{dc^1(t)}{dt} = -\gamma(\mu_1+\psi_1)c^1(t) - \gamma^2\psi_1^2\left(1 - \frac{\psi_1}{\mu_1+\psi_1}(1-e^{-\gamma(\mu_1+\psi_1)t})\right)\int_0^t f(t-s)e^{-\gamma(\mu_1+\psi_1)s}ds. \quad (23)$$

Again, (23) is a first-order ordinary differential equation which can be solved to obtain

$$c^1(t) = -\frac{\gamma^2\psi_1^3}{\mu_1+\psi_1}e^{-\gamma(\mu_1+\psi_1)t}\int_0^t\int_0^u\left(\frac{\mu_1}{\psi_1}e^{\gamma(\mu_1+\psi_1)u}+1\right)f(u-s)e^{-\gamma(\mu_1+\psi_1)s}dsdu. \quad (24)$$

The double integral in Eq. (24) can be simplified as follows by changing the order of integration

$$\int_0^t\int_s^t\left(\frac{\mu_1}{\psi_1}e^{\gamma(\mu_1+\psi_1)u}+1\right)f(u-s)e^{-\gamma(\mu_1+\psi_1)s}duds.$$

Now apply the transformation of $u = v + s$ and changing the order of integration again, yields

$$\frac{1}{\gamma(\mu_1+\psi_1)}\int_0^t\left(\frac{\gamma(\mu_1+\psi_1)\mu_1}{\psi_1}(t-v)e^{\gamma(\mu_1+\psi_1)v}+1-e^{-\gamma(\mu_1+\psi_1)(t-v)}\right)f(v)dv.$$

Therefore,

$$c^1(t) = -\frac{\gamma\psi_1^3}{(\mu_1+\psi_1)^2}e^{-\gamma(\mu_1+\psi_1)t}\times\int_0^t\left(\frac{\gamma(\mu_1+\psi_1)\mu_1}{\psi_1}(t-v)e^{\gamma(\mu_1+\psi_1)v}+1-e^{-\gamma(\mu_1+\psi_1)(t-v)}\right)f(v)dv,$$

which gives an approximation of the function $c(t)$ that can be evaluated numerically.

3 The Competitive Assay

Competitive binding immunoassays are based on antibody-antigen interactions in which the number of antigen binding sites on the antibody is limited. The antigen and a labelled analogue are incubated together with a fixed concentration of the antibody and the signal produced will reflect the competition between the antigen and analogue for binding to the antibody. This method requires that the antibody should have the same binding affinity for the antigen as for the labelled analogue; we also assume that the probability of binding to antibody is the same for both species.

3.1 Simplified Model for the Competitive Assay

As in the previous section, we start by studying the kinetics of the chemical reactions in a competitive assay in the absence of any transport effects. The antibody-antigen interactions with competition can be expressed symbolically as

$$A + B \underset{k_-}{\overset{k}{\rightleftarrows}} C, \qquad A^{'} + B \underset{k_-}{\overset{k}{\rightleftarrows}} C^{'}, \tag{25}$$

where A, B and C are the same as defined in Sect. 2.1; $A^{'}$ is basically the antigen with a label attached to it called an *analogue*, and $C^{'}$ is the product formed by the antibody and analogue. We assume the two reactions have the same forward and backward rate constants of k and k_-, and we denote the concentration of all reactants and products by their corresponding lower case letters, namely

$$a = [A], \quad a^{'} = [A^{'}], \quad b = [B], \quad c = [C], \quad c^{'} = [C^{'}].$$

The dynamics of the system is described by the following system of dimensionless equations

$$\begin{cases} \frac{da}{dt} = -ab + \mu c & \text{(a)} \\ \frac{da^{'}}{dt} = -a^{'}b + \mu c^{'} & \text{(b)} \\ \frac{db}{dt} = -b(a + a^{'}) + \mu(c + c^{'}) & \text{(c)} \\ \frac{dc}{dt} = ab - \mu c & \text{(d)} \\ \frac{dc^{'}}{dt} = a^{'}b - \mu c^{'}. & \text{(e)} \end{cases} \tag{26}$$

The non-dimensional initial conditions are:

$$a(0) = \psi, \quad a^{'}(0) = \psi^{'}, \quad b(0) = 1, \quad c(0) = 0, \quad c^{'}(0) = 0,$$

and the non-dimensional conservation laws become

$$\begin{cases} a + c = \psi & \text{(a)} \\ a^{'} + c^{'} = \psi^{'} & \text{(b)} \\ b + c + c^{'} = 1, & \text{(c)} \end{cases} \tag{27}$$

where we define

$$\psi = \frac{a_0}{b_0}, \quad \psi^{'} = \frac{a_0^{'}}{b_0}, \quad \mu = \frac{k_-}{kb_0}. \tag{28}$$

Using similar calculations to those shown in the previous section, system (26) together with the conservation laws (27) yield the following equilibrium equation for the antibody concentration, b,

$$b^2 - \left(1 - \psi - \psi^{'} - \mu\right) b - \mu = 0. \tag{29}$$

The exact values of the steady states for all the species are as follows:

$$\begin{cases} b = \frac{1}{2}\left(1 - \psi - \psi^{'} - \mu + \sqrt{(1 - \psi - \psi^{'} - \mu)^2 + 4\mu}\right) & \text{(a)} \\ a = \frac{\psi\mu}{\mu+b} & \text{(b)} \\ a^{'} = \frac{\psi^{'}\mu}{\mu+b} & \text{(c)} \\ c = \frac{\psi b}{\mu+b} & \text{(d)} \\ c^{'} = \frac{\psi^{'} b}{\mu+b}. & \text{(e)} \end{cases} \qquad (30)$$

We are now going to calculate the asymptotic approximations to these solutions as shown in (30), in a manner similar to the previous model. Again we start with an expansion of the form

$$b = \tilde{b_0} + \mu b_1 + \mu^2 b_2 + \cdots \qquad (31)$$

(we have used the notation $\tilde{b_0}$ for the first term of the expansion in order to avoid confusing it with b_0, the initial antibody concentration). Then substituting Eq. (31) into (29), we get

$$(\tilde{b_0} + \mu b_1 + \mu^2 b_2 + \cdots)^2 - \left(1 - \psi - \psi^{'} - \mu\right)(\tilde{b_0} + \mu b_1 + \mu^2 b_2 + \cdots) - \mu = 0.$$

By collecting coefficients of powers of μ, at $O(1)$, $O(\mu)$ and $O(\mu^2)$ we obtain

$$b = \mu\frac{1}{\psi + \psi^{'} - 1} - \mu^2\frac{\psi + \psi^{'}}{\left(\psi + \psi^{'} - 1\right)^3} + \cdots, \quad \text{if} \quad \psi + \psi^{'} > 1,$$

or

$$b = 1 - \psi - \psi^{'} + \mu\frac{\psi + \psi^{'}}{1 - \psi - \psi^{'}} + \cdots, \quad \text{if} \quad \psi + \psi^{'} < 1.$$

In the case where $\psi + \psi^{'} = 1$, we cannot choose either of the two solutions, since the denominators in the solutions are equal to zero. In this case, we have to start the asymptotic analysis again with $\psi + \psi^{'} = 1$ substituted into Eq. (29). Thus the equilibrium values for b are given by the equation

$$b^2 + \mu b - \mu = 0. \qquad (32)$$

Again, in this case it is more appropriate to use the expansion

$$b = \tilde{b_0} + \sqrt{\mu} b_1 + \mu b_2 + \mu\sqrt{\mu} b_3 + \cdots \qquad (33)$$

since we can clearly see that there is an $\sqrt{\mu}$ term contained in Eq. (30a). Substituting the new expansion (33) into Eq. (32), we obtain the equation

$$(\tilde{b_0} + \sqrt{\mu} b_1 + \mu b_2 + \mu\sqrt{\mu} b_3 + \cdots)^2 + \mu(\tilde{b_0} + \sqrt{\mu} b_1 + \mu b_2 + \mu\sqrt{\mu} b_3 + \cdots) - \mu = 0,$$

and then by collecting coefficients of powers of $\sqrt{\mu}$, at $O(1)$, $O(\sqrt{\mu})$, $O(\mu)$ and $O(\mu\sqrt{\mu})$ yields

$$b = \sqrt{\mu} - \frac{1}{2}\mu + \cdots$$

or

$$b = -\sqrt{\mu} - \frac{1}{2}\mu + \cdots, \text{ which cannot be a solution, since } b > 0.$$

Now we need to find the solutions for a, $a^{'}$, c and $c^{'}$. From Eq. (26a) and (27a), we get

$$\frac{da}{dt} = -ab + \mu(\psi - a),$$

which indicates that the equilibrium value of a can be obtained. Here, we are going to use the same asymptotic expansion (33) for b, and use the expansion for a as

$$a = \tilde{a_0} + \mu a_1 + \mu^2 a_2 + \cdots \tag{34}$$

To summarise, the equilibrium solutions for the reactants and products in the three cases discussed above are as follows:

Case 1: When $b_0 < a_0 + a_0^{'}$ (i.e., $\psi + \psi^{'} > 1$), the equilibrium solutions are

$$\begin{cases} a = \psi - \frac{\psi}{\psi+\psi^{'}} + \mu\frac{\psi}{(\psi+\psi^{'}-1)(\psi+\psi^{'})} + \cdots & \text{(a)} \\ a^{'} = \psi^{'} - \frac{\psi^{'}}{\psi+\psi^{'}} + \mu\frac{\psi^{'}}{(\psi+\psi^{'}-1)(\psi+\psi^{'})} + \cdots & \text{(b)} \\ b = \frac{\mu}{\psi+\psi^{'}-1} - \mu^2\frac{\psi+\psi^{'}}{(\psi+\psi^{'}-1)^3} + \cdots & \text{(c)} \\ c = \frac{\psi}{\psi+\psi^{'}} - \mu\frac{\psi}{(\psi+\psi^{'}-1)(\psi+\psi^{'})} + \cdots & \text{(d)} \\ c^{'} = \frac{\psi^{'}}{\psi+\psi^{'}} - \mu\frac{\psi^{'}}{(\psi+\psi^{'}-1)(\psi+\psi^{'})} + \cdots & \text{(e)} \end{cases} \tag{35}$$

Case 2: When $b_0 > a_0 + a_0^{'}$ (i.e., $\psi + \psi^{'} < 1$), the equilibrium solutions are

$$\begin{cases} a = \mu\frac{\psi}{1-\psi-\psi^{'}} + \cdots & \text{(a)} \\ a^{'} = \mu\frac{\psi^{'}}{1-\psi-\psi^{'}} + \cdots & \text{(b)} \\ b = 1 - \psi - \psi^{'} + \mu\frac{\psi+\psi^{'}}{1-\psi-\psi^{'}} + \cdots & \text{(c)} \\ c = \psi - \mu\frac{\psi}{1-\psi-\psi^{'}} + \cdots & \text{(d)} \\ c^{'} = \psi^{'} - \mu\frac{\psi^{'}}{1-\psi-\psi^{'}} + \cdots & \text{(e)} \end{cases} \tag{36}$$

Case 3: When $b_0 = a_0 + a_0^{'}$ (i.e., $\psi + \psi^{'} = 1$), the equilibrium solutions are

$$\begin{cases} a = \sqrt{\mu}\psi - \frac{1}{2}\mu\psi + \cdots & \text{(a)} \\ a^{'} = \sqrt{\mu}\psi^{'} - \frac{1}{2}\mu\psi^{'} + \cdots & \text{(b)} \\ b = \sqrt{\mu} - \frac{1}{2}\mu + \cdots & \text{(c)} \\ c = \psi - \sqrt{\mu}\psi + \frac{1}{2}\mu\psi + \cdots & \text{(d)} \\ c^{'} = \psi^{'} - \sqrt{\mu}\psi^{'} + \frac{1}{2}\mu\psi^{'} + \cdots & \text{(e)} \end{cases} \qquad (37)$$

Note that if $a_0^{'} = 0$ (i.e., labelled antigen is absent), the assay is no longer a competition system and solutions (35)–(37) reduce to the solutions (11)–(13) obtained in Sect. 2.1. Also note that the behaviour of the competitive system is qualitatively different in the three cases discussed above. Case 1 ($b_0 < a_0 + a_0^{'}$) is the case which is most relevant to experiments, since antibody sites are limited so there is a true competition between antigen and analogue. In this case, the equilibrium solutions show that antibody sites are almost depleted while antigen and analogue bind to a ratio equal to that of their initial concentrations. In Case 2 ($b_0 > a_0 + a_0^{'}$), the antibody binding sites are plentiful and so all antigen and analogue molecules will eventually bind and form products.

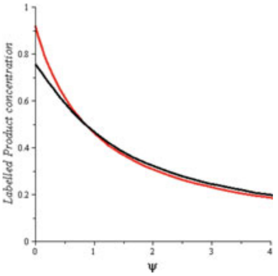
We now show how these results can be used for constructing calibration curves for competitive systems. The solutions of antibody-antigen interactions with competition model were considered in the case of $b_0 < a_0 + a_0^{'}$, $b_0 > a_0 + a_0^{'}$ and $b_0 = a_0 + a_0^{'}$. In a real-life testing situation, a_0 is unknown, so the analysis below is more appropriate (we assume that b_0 and $a_0^{'}$ are given).

Case I: When $b_0 \leq a_0^{'}$ (i.e., $\psi^{'} > 1$); this implies $b_0 < a_0 + a_0^{'}$ or $\psi + \psi^{'} > 1$, since a_0 is positive. The solution in this case is identical to the solution obtained in Eq. (35) presented above. The expression of the labelled product in terms of ψ and $\psi^{'}$ is given by

$$c^{'} = \frac{\psi^{'}}{\psi + \psi^{'}} - \mu\frac{\psi^{'}}{(\psi + \psi^{'} - 1)(\psi + \psi^{'})} + \cdots$$

We plot $c^{'}$ against ψ, as given by the proceeding formula, to get the calibration curve (red) in Fig. 2. This is compared with the plot of the exact solution (black) given by Eq. (30e) and the two curves are in good agreement for $\mu \ll 1$. Note that, since

Fig. 2 Exact value (black) and asymptotic approximation (red) for the labelled product as functions of ψ in Case I. Typical values for constants used in this simulation are: $b_0 = 1$, $a_0^{'} = 1$, $k = 100$ and $k_{-} = 8$



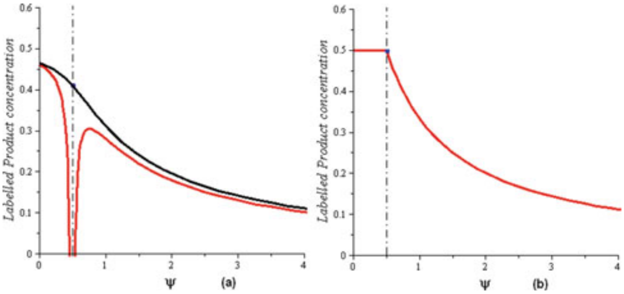


Fig. 3 Exact value (black) and asymptotic approximation (red) for the labelled product as functions of ψ in Case II. Typical values for constants used in this simulation are: $b_0 = 2$, $a_0^{'} = 1$, $k = 100$, $k_- = 8$ in (a) and $k_- = 0$ in (b)

$\psi^{'} > 1$, the asymptotic approximation in this case is uniformly valid for all values of $\psi > 0$.

Case II: When $b_0 > a_0^{'}$ (i.e., $\psi^{'} < 1$), we need to consider the following two situations;

- If $b_0 > a_0 + a_0^{'}$ then $\psi < 1 - \psi^{'}$, and the solution for $c^{'}$ is given by Eq. (36e) (Case 2 in the previous analysis);
- If $b_0 < a_0 + a_0^{'}$ then $\psi > 1 - \psi^{'}$, and the solution for $c^{'}$ is given by Eq. (35e) (Case 1 in the previous analysis).

Therefore, we conclude the solution for $c^{'}$ is given by

$$c^{'} = \begin{cases} \frac{\psi^{'}}{\psi+\psi^{'}} - \mu \frac{\psi^{'}}{(\psi+\psi^{'}-1)(\psi+\psi^{'})} + \cdots, & \text{if } \psi + \psi^{'} > 1 \ (b_0 < a_0 + a_0^{'}) \\ \psi^{'} - \mu \frac{\psi^{'}}{1-\psi-\psi^{'}} + \cdots, & \text{if } \psi + \psi^{'} < 1 \ (b_0 > a_0 + a_0^{'}) \\ \psi^{'} - \sqrt{\mu}\psi^{'} + \frac{1}{2}\mu\psi^{'} + \cdots, & \text{if } \psi + \psi^{'} = 1 \ (b_0 = a_0 + a_0^{'}). \end{cases}$$

Combining these three solution branches, we obtain the plots shown in Fig. 3, which are shown together with the exact solution for $c^{'}$ given by Eq. (30e).

We make the same remark as in the case of direct assays, namely that the asymptotic approximation for $c^{'}$ in this case is not uniformly valid around $\psi = 1 - \psi^{'}$. Once again, the restriction $\psi + \psi^{'} > 1$ applies in most practical situations so that non-uniformity will not be relevant in this region. Note also that, when $\mu = 0$, the asymptotic approximation is identical to equations in system (30).

3.2 *Diffusion Model for the Competitive Assay*

We now consider the case when some of the reactants are free to diffuse within a small cell, modelled as a one-dimensional domain. Just as in Sect. 2.2, we assume that the antibody is immobilised to a surface (in our one-dimensional model this actually corresponds to one point) while the antigen and labelled antigen can move throughout the cell. A consistent non-dimensional system of equations which describes the behaviour of the relevant chemical species is given by

$$\begin{cases} \frac{\partial a(x,t)}{\partial t} = \frac{\partial^2 a(x,t)}{\partial x^2} & \text{(a)} \\ \frac{\partial a^{'}(x,t)}{\partial t} = \frac{\partial^2 a^{'}(x,t)}{\partial x^2} & \text{(b)} \\ a(x,0) = \psi_1 & \text{(c)} \\ a^{'}(x,0) = \psi_2 & \text{(d)} \\ \frac{\partial a(0,t)}{\partial x} = 0 & \text{(e)} \\ \frac{\partial a^{'}(0,t)}{\partial x} = 0 & \text{(f)} \\ \frac{\partial a(1,t)}{\partial x} = \gamma\left(\mu_1 c(t) - a(1,t)\left(1 - c(t) - c^{'}(t)\right)\right) & \text{(g)} \\ \frac{\partial a^{'}(1,t)}{\partial x} = \gamma\left(\mu_1 c^{'}(t) - a^{'}(1,t)\left(1 - c(t) - c^{'}(t)\right)\right) & \text{(h)} \\ c(t) + \int_0^1 a(x,t)dx = \psi_1 & \text{(i)} \\ c'(t) + \int_0^1 a^{'}(x,t)dx = \psi_2, & \text{(j)} \end{cases} \tag{38}$$

with $x \in (0, 1)$, $t \geq 0$, and we define

$$\psi_1 = \frac{a_0 d}{b_0}, \quad \gamma = \frac{dkb_0}{D}, \quad \mu_1 = \frac{k\_d}{kb_0}, \quad \psi_2 = \frac{a_0^{'}d}{b_0}.$$

Next, we are going to analyse system (38) as $t \to \infty$. At equilibrium,

$$b^* = \frac{1}{2}\left(1 - \mu_1 - \psi_1 - \psi_2 + \sqrt{(1 - \mu_1 - \psi_1 - \psi_2)^2 + 4\mu_1}\right). \tag{39}$$

Also, we obtain the solutions for $a^*(x)$, $a^{'*}(x)$, c^* and $c^{'*}$ as

$$\begin{cases} a^* = \frac{\psi_1 \mu_1}{\mu_1 + b^*} & \text{(a)} \\ a^{'*} = \frac{\psi_2 \mu_1}{\mu_1 + b^*} & \text{(b)} \\ c^* = \frac{\psi_1 b^*}{\mu_1 + b^*} & \text{(c)} \\ c^{'*} = \frac{\psi_2 b^*}{\mu_1 + b^*}, & \text{(d)} \end{cases} \tag{40}$$

which are the same solutions as shown in (30) obtained in Sect. 3.1 (since d/b_0 in the diffusion model is equivalent to $1/b_0$ in the non-diffusion model). Using Laplace transforms and their properties we carry out a similar calculation to that given in Sect. 2.2 and obtain the following system of Volterra integro-differential equations:

$$\begin{cases} \dfrac{dc(t)}{dt} = \gamma\psi_1 - \gamma(\mu_1+\psi_1)c(t) - \gamma\psi_1 c^{'}(t) \\ \qquad -\gamma\left(1-c(t)-c^{'}(t)\right)\displaystyle\int_0^t f(t-s)\frac{dc}{ds}(s)ds & \text{(a)} \\ \dfrac{dc^{'}(t)}{dt} = \gamma\psi_2 - \gamma\psi_2 c(t) - \gamma\left(\mu_1+\psi_2\right)c^{'}(t) \\ \qquad -\gamma\left(1-c(t)-c^{'}(t)\right)\displaystyle\int_0^t f(t-s)\frac{dc^{'}}{ds}(s)ds, & \text{(b)} \end{cases} \qquad (41)$$

where $f(t)$ has the same definition as in Sect. 2.2. (see (21)). Note that in the absence of labelled antigen ($c^{'} = 0$), Eq. (41a) yields the result obtained for the non-competitive assay (see (18)).

Adding (41a) and (41b) and, if we use the conservation law $c(t) + c^{'}(t) = 1 - b(t)$, we get

$$\frac{db(t)}{dt} = \gamma\mu_1 - \gamma(\mu_1+\psi_1+\psi_2)b(t) - \gamma b(t)\int_0^t f(t-s)\frac{db}{ds}(s)ds. \qquad (42)$$

Once the solution of $b(t)$ is calculated (using, for example, the asymptotic or numerical methods detailed in [1, 2]), the product concentration $c^{'}(t)$ can be determined from Eq. (41b), which yields

$$\frac{dc^{'}(t)}{dt} = -\gamma\mu_1 c^{'}(t) + \gamma b(t)\left(\psi_2 - \int_0^t f(t-s)\frac{dc^{'}}{ds}(s)ds\right). \qquad (43)$$

Hence, instead of solving a coupled system of integro-differential Eq. (41), we can now solve the independent Eq. (42) followed by Eq. (43). A regular perturbation analysis could be applied to (42) and (43), which is similar to the one used in Sect. 2.2 to obtain an approximation for $b(t)$ and $c^{'}(t)$.

We conclude that, our assumption of identical rate constants for antibody-antigen and antibody-analogue binding leads to a significant simplification of the problem studied in [5], whereby a coupled system of Volterra integro-differential equations was replaced by an uncoupled one. However, this simplification may not always be feasible since the label attachment may interfere with the antigen's epitope and therefore has to be considered carefully for each experimental setting.

4 The Sandwich Assay

The Sandwich assay (refer to, for example, [6]) is a type of immunoassay in which an antibody for the antigen to be assayed is immobilised to a solid surface (this antibody is often referred to as the capture antibody), then the sample containing the test analyte is added and the reaction has been allowed to reach equilibrium.

A second antibody, which has a radioactive or fluorescent label (and is therefore called a tracer) is added, sandwiching the antigen. Again, after removal of excess, the amount of bound label is measured. The signal level in this type of assay is clearly proportional to the analyte concentration in the sample, just like in the direct assay. The second antibody may be specific for a different epitope on the antigen, thus enhancing overall specificity, or for the first antibody bound to an antigen. This process can be symbolically represented by the reactions given by (44).

$$A + B_1 \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C_1, \quad D + B_1 \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C_2, \quad A + B_2 \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} D, \quad C_1 + B_2 \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} C_2, \tag{44}$$

where A represents antigen, B_1 represents the immobilised antibody (also referred to as capture antibody), B_2 represents the labelled antibody, C_1 is the product of the antigen and immobilised antibody, C_2 is the product of C_1 and labelled antibody (also referred to as the sandwich product), and D is the product of antigen and labelled antibody. The first two reactions have a forward reaction rate of k_1 and a backward reaction rate of k_{-1}, the third and fourth reactions have a forward and backward reaction rate of k_2 and k_{-2} respectively. We have assumed that the affinity of each antibody for the corresponding antigen is the same regardless of whether the antigen is free or bound to another antibody; this simplifying assumption is not essential for the model and could easily be relaxed later. We denote the concentration of the reactants and products by their corresponding lower case letters, i.e.,

$$a = [A], \quad b_1 = [B_1], \quad b_2 = [B_2], \quad c_1 = [C_1], \quad c_2 = [C_2], \quad d = [D].$$

The dynamic of the system is described by the following non-dimensional system

$$\begin{cases} \frac{da}{dt} = K_{-1}c_1 + K_{-2}d - ab_1 - K^{'}ab_2 & \text{(a)} \\ \frac{db_1}{dt} = K_{-1}(c_1 + c_2) - b_1(a + d) & \text{(b)} \\ \frac{db_2}{dt} = K_{-2}(c_2 + d) - K^{'}b_2(a + c_1) & \text{(c)} \\ \frac{dc_1}{dt} = ab_1 - K_{-1}c_1 - K^{'}b_2c_1 + K_{-2}c_2 & \text{(d)} \\ \frac{dc_2}{dt} = K^{'}b_2c_1 + b_1d - (K_{-1} + K_{-2})c_2 & \text{(e)} \\ \frac{dd}{dt} = K^{'}ab_2 - K_{-2}d - b_1d + K_{-1}c_2, & \text{(f)} \end{cases} \tag{45}$$

where we let

$$K_{-1} = \frac{k_{-1}}{k_1\beta_1}, \quad K_{-2} = \frac{k_{-2}}{k_1\beta_1}, \quad K^{'} = \frac{k_2}{k_1}. \tag{46}$$

The non-dimensional initial conditions and conservation laws are:

$$a(0) = \frac{\alpha}{\beta_1}, \quad b_1(0) = 1, \quad b_2(0) = \frac{\beta_2}{\beta_1}, \quad c_1(0) = 0, \quad c_1(0) = 0, \quad d(0) = 0, \tag{47}$$

and

$$\begin{cases} a + c_1 + c_2 + d = \frac{\alpha}{\beta_1} & \text{(a)} \\ b_1 + c_1 + c_2 = 1 & \text{(b)} \\ b_2 + c_2 + d = \frac{\beta_2}{\beta_1}. & \text{(c)} \end{cases} \tag{48}$$

From the steady state forms of Eq. (45) and the conservation laws (48) we find that

$$b_1^2 + \left(\frac{\alpha}{\beta_1} - 1 + K_{-1}\right) b_1 - K_{-1} = 0, \quad b_2^2 + \left(\frac{\alpha - \beta_2}{\beta_1} + \frac{K_{-2}}{K'}\right) b_2 - \frac{K_{-2}\beta_2}{K'\beta_1} = 0,$$

and

$$c_2 = \frac{\frac{\beta_2}{\beta_1} b_1 + K^{'} b_2 - (1 + K^{'}) b_1 b_2}{b_1 + K_{-1} + K^{'} b_2 + K_{-3}}, \quad K_{-3} = \frac{k_{-3}}{k_1 \beta_1}.$$

Therefore, it is possible to calculate the exact values of the steady states for all the species, provided that all the reaction constants and initial concentrations are accurately known. Some calibration curves, consisting of the steady states of c_2, $c_2 + d$ and b_2 as functions of initial antigen concentration, α are plotted in Fig. 4. The reason for plotting these species is that some antibodies have radioactive or fluorescent labels which can be measured both at the surface and in the solution. If the signal is measured at the surface, we need to plot c_2 and compare it with experimental data; however, for signals measured in the solution, it is $c_2 + d$ we are interested in. Note also that over the short initial stage of the reaction, there exists a linear response between the signal and analyte concentration.

The performance of a biosensor is often affected by the presence of a **non-specific** (noisy) component of the recorded signal. In the configuration described above, any measurement of the fluorescent label in solution would inevitably include B_2, which is the amount of labelled antibody left over (or unbound) after the reaction. This is a non-specific measurement as it does not provide any information about the antigen in the sample. We have also plotted b_2, the noise, together with the "good" signals in Fig. 4.

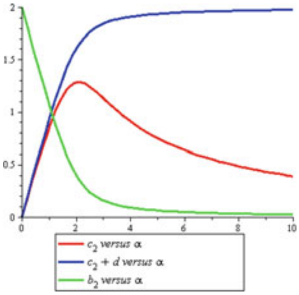


Fig. 4 Sandwich product c_2 (red), combined product $c_2 + d$ (blue), and unbound tracer b_2 (green) as functions of initial antigen concentration α. Typical values for constants used in this simulation are: $k_1 = 100$, $k_{-1} = 10$, $k_2 = 100$, $k_{-2} = 10$, $\beta_1 = 2$ and $\beta_2 = 2$

An alternative modelling strategy is to construct a two-step model. In the first step, we add antigen to the capture antibody and allow the reaction

$$A + B_1 \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C_1, \tag{49}$$

to proceed to equilibrium. This corresponds to the direct assay model studied in Sect. 2.1, where exact and approximate formulas were obtained for the equilibrium value of C_1. After the unbound antigen is washed away, we construct a second model where the labelled antibody is introduced in the system (which does not contain any free antigen) and reacts with the product C_1 to form C_2,

$$C_1 + B_2 \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} C_2. \tag{50}$$

The equilibrium value of c_2 can then be obtained as a function of c_1 and hence of the initial antigen concentration, a_0. Note that this modelling strategy does not eliminate noise completely as, even after washing, the reversible nature of the reactions (49) and (50) means that small amounts of free A and C_1 will still be present in the solution. (However, in an experimental setting, washing is always practiced since it greatly reduces these amounts hence minimising non-specific interactions).

The analysis of this two-step model is similar to the one presented above and will not be given here; instead, it will be performed as part of future studies into sandwich bioassays (refer to the conclusions section). What this example illustrates is how, in a simple model, it is possible to distinguish between the specific signal and the noise and we believe that this calculation should bring valuable insight into experimental procedures.

5 Summary

In this paper we analysed several modelling strategies for antibody-antigen interactions with possible applications to immunoassay design. For direct and competitive assays, we constructed two types of mathematical models: one-point models which describe only the reaction kinetics and spatially extended models which allowed for transport of one or more species to the reaction site. For both these assays (and both types of models) we were able to derive exact formulas for the equilibrium values of all reactants and construct calibration curves, which give the final product as a function of the initial analyte concentration. It was concluded that, for each of the assays considered, both modelling approaches gave identical equilibrium values and hence the biosensor response was the same regardless of whether transport effects were included in the model or not. Therefore, if the value of the equilibrium state is the only piece of information required in an experimental context we would recommend using the simpler model without diffusion. However, in many practical problems, the time taken to achieve equilibrium is also a parameter of interest and, in such cases, we obviously cannot neglect transport. Our modelling results were found to agree with the results in [1, 2] and [5] which presented more detailed and rigorous mathematical studies of diffusion models for similar direct and competitive assays. As remarked before, the aim of our work in this chapter was to find conditions under which simpler models and studies could be used in the context of antibody-antigen interactions. The last section presented a different type of immunoassay, namely a sandwich assay, for which a simple one-point model was used in order to construct a calibration curve. This example illustrated how mathematical modelling has the potential to evaluate the ratio between specific and non-specific signals in an experimental problem and optimise biosensor performance by identifying parameter regions where the noise is minimal.

References

1. Jones, S., Jumarhon, B., McKee, S., Scott, J.A.: A mathematical model of a biosensor. J. Eng. Math. **30**, 321–337 (1996)
2. Jumarhon, B., McKee, S.: On the heat equation with nonlinear and nonlocal boundary conditions. J. Math. Anal. Appl. **190**, 806–820 (1995)
3. Badley, R.A., Drake, R.A.L., Shanks, I.A., Smith, A.M., Stephenson, P.R.: Optical biosensors for immunoassays, the fluorescence capillary-fill device. Philos. Trans. R. Soc. Lond. Ser. B **316**, 143–160 (1987)
4. Debnath, L.: Integral Transforms and Their Applications. CRC Press, Boca Raton (1995)
5. De Jesus Rebelo, M.S.: Analytical and numerical methods for nonlinear volterra integral equations with weakly singular kernel. Ph.D thesis (2009)
6. Wild, D.: The Immunoassay Handbook. Elsevier, Amsterdam (2005)

Positive Solutions for a Class of Nonlocal Discrete Boundary Value Problems

Rodica Luca

Abstract We study the existence of positive solutions for a nonlinear second-order difference equation with a linear term, a parameter and a sign-changing nonlinearity, subject to multi-point boundary conditions. In the proof of our main results we use the Guo-Krasnosel'skii fixed point theorem and the nonlinear alternative of Leray-Schauder type.

1 Introduction

We consider the nonlinear second-order difference equation

$$(E) \qquad \Delta^2 u_{n-1} - Lu_n + \lambda f(n, u_n) = 0, \quad n = \overline{1, N-1},$$

with the multi-point boundary conditions

$$(BC) \qquad u_0 = \sum_{i=1}^{p} a_i u_{\xi_i}, \quad u_N = \sum_{i=1}^{q} b_i u_{\eta_i},$$

where $N \in \mathbf{N}$, $N > 2$, $p,\ q \in \mathbf{N}$, Δ is the forward difference operator with stepsize 1, $\Delta u_n = u_{n+1} - u_n$, $\Delta^2 u_{n-1} = u_{n+1} - 2u_n + u_{n-1}$, and $n = \overline{k, m}$ means that $n = k,\ k+1, \ldots, m$ for $k,\ m \in \mathbf{N}$, $a_i \in \mathbf{R}$ and $\xi_i \in \mathbf{N}$ for all $i = \overline{1, p}$, $b_i \in \mathbf{R}$ and $\eta_i \in \mathbf{N}$ for all $i = \overline{1, q}$, $1 \le \xi_1 < \cdots < \xi_p \le N-1$, $1 \le \eta_1 < \cdots < \eta_q \le N-1$, L is a positive constant, λ is a positive parameter and f is a sign-changing nonlinearity.

Under some assumptions on the nonlinearity f, we present intervals for the parameter λ such that problem $(E) - (BC)$ has positive solutions $(u_n)_{n=\overline{0,N}}$ with $u_n > 0$ for all $n = \overline{0, N}$. The problem $(E) - (BC)$ with $L = 0$ and $a_i = 0$ for all $i = \overline{1, p}$ was recently studied in the paper [20]. The existence of positive solutions for the

R. Luca (✉)
Department of Mathematics, Gh. Asachi Technical University,
11 Blvd. Carol I, 700506 Iasi, Romania
e-mail: rluca@math.tuiasi.ro; rlucatudor@yahoo.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_46

problem $(E)-(BC)$ with $\lambda=1$ (that is, without parameter) was investigated in [21] by using the Guo-Krasnosel'skii fixed point theorem. The equation (E) with $\lambda=1$ and $L=0$, where the nonlinearity f may be unbounded below or non-positive, subject to the boundary conditions $u_0=u_1$ and $u_N=u_{N-1}$, which is a resonant problem, has been studied in the paper [8] by transforming it into a non-resonant problem. The existence, nonexistence and multiplicity of positive solutions for difference equations and systems of difference equations with parameters or without parameters, with non-negative or sign-changing nonlinearities, supplemented with various boundary conditions were investigated in the papers [1, 3, 5–7, 9–11, 13, 15–17, 22–24] and the monograph [14]. We also recommend the readers the monographs [2, 18] and [19] for various applications of the nonlinear difference equations in many domains.

2 Preliminary Results

In this section we present some auxiliary results from [21] that will be used in the proof of our main theorems.

We consider the second-order difference equation

$$\Delta^2 u_{n-1} - Lu_n + y_n = 0, \quad n=\overline{1,N-1}, \tag{1}$$

with the multi-point boundary conditions (BC), where $y_n \in \mathbf{R}$ for all $n=\overline{1,N-1}$. We denote by $A=\frac{L+2+\sqrt{L^2+4L}}{2}$ the biggest solution of the characteric equation $r^2-(L+2)r+1=0$ associated to Eq. (1). We also denote by

$$\begin{aligned}\Delta_1 = &\left(\sum_{i=1}^{p} a_i A^{\xi_i} - 1\right)\left(\frac{1}{A^N} - \sum_{i=1}^{q} b_i \frac{1}{A^{\eta_i}}\right)\\ &+\left(1-\sum_{i=1}^{p} a_i \frac{1}{A^{\xi_i}}\right)\left(A^N - \sum_{i=1}^{q} b_i A^{\eta_i}\right).\end{aligned}$$

Lemma 1 ([21]). *If $\Delta_1 \neq 0$, then the unique solution of problem (1)-(BC) can be expressed as $u_n=\sum_{j=1}^{N-1} G(n,j)y_j$, $n=\overline{0,N}$, where the Green function G is given by*

$$\begin{aligned}G(n,j) &= g(n,j)\\ &+\frac{1}{\Delta_1}\left[A^n\left(\sum_{i=1}^{q} b_i \frac{1}{A^{\eta_i}} - \frac{1}{A^N}\right) + \frac{1}{A^n}\left(A^N - \sum_{i=1}^{q} b_i A^{\eta_i}\right)\right]\sum_{i=1}^{p} a_i g(\xi_i,j)\\ &+\frac{1}{\Delta_1}\left[A^n\left(1-\sum_{i=1}^{p} a_i \frac{1}{A^{\xi_i}}\right) + \frac{1}{A^n}\left(\sum_{i=1}^{p} a_i A^{\xi_i} - 1\right)\right]\sum_{i=1}^{q} b_i g(\eta_i,j),\end{aligned} \tag{2}$$

for all $n = \overline{0, N}$ *and* $j = \overline{1, N-1}$*, and the function* g *is given by*

$$g(n, j) = \frac{A}{(A^2-1)(A^N - A^{-N})} \times \begin{cases} (A^j - A^{-j})(A^{N-n} - A^{n-N}), & 1 \le j < n \le N, \\ (A^n - A^{-n})(A^{N-j} - A^{j-N}), & 0 \le n \le j \le N-1. \end{cases}$$

Lemma 2 ([21]). *We assume that* $a_i \ge 0$ *for all* $i = \overline{1, p}$*,* $b_j \ge 0$ *for all* $j = \overline{1, q}$*,* $\sum_{i=1}^p a_i A^{\xi_i} \ge 1$*,* $\sum_{i=1}^p a_i \frac{1}{A^{\xi_i}} \le 1$*,* $\sum_{i=1}^q b_i \frac{1}{A^{\eta_i}} \ge \frac{1}{A^N}$*,* $\sum_{i=1}^q b_i A^{\eta_i} \le A^N$*, and* $\Delta_1 > 0$*. Then the Green function* G *given by* (2) *satisfies the inequalities*

$$k(n)h(j) \le G(n, j) \le \Lambda h(j), \quad \forall n = \overline{0, N}, \ \ j = \overline{1, N-1},$$

where

$$\Lambda = 1 + \frac{1}{\Delta_1}\left[A^N\left(\sum_{i=1}^q b_i \frac{1}{A^{\eta_i}} - \frac{1}{A^N}\right) + A^N - \sum_{i=1}^q b_i A^{\eta_i}\right]\left(\sum_{i=1}^p a_i\right)$$
$$+\frac{1}{\Delta_1}\left[A^N\left(1 - \sum_{i=1}^p a_i \frac{1}{A^{\xi_i}}\right) + \sum_{i=1}^p a_i A^{\xi_i} - 1\right]\left(\sum_{i=1}^q b_i\right),$$
$$h(j) = \frac{A}{(A^2-1)(A^N-A^{-N})}(A^j - A^{-j})(A^{N-j} - A^{j-N}), \quad \forall j = \overline{1, N-1},$$
$$k(n) = \frac{1}{A^{N-1} - A^{1-N}} \min\{A^n - A^{-n}, A^{N-n} - A^{n-N}\}, \quad \forall n = \overline{0, N}.$$

Lemma 3 ([21]). *Under the assumptions of Lemma* 2*, the solution* u_n*,* $n = \overline{0, N}$ *of problem* (1)*-*(BC) *satisfies the inequality* $u_n \ge \frac{1}{\Lambda} k(n) u_m$ *for all* $n, m = \overline{0, N}$*.*

We also present here the Guo-Krasnosel'skii fixed point theorem (see [12]) and the nonlinear alternative of Leray-Schauder type (see [4]) that will be used in the next section.

Theorem 1. *Let* X *be a Banach space and let* $C \subset X$ *be a cone in* X*. Assume* Ω_1 *and* Ω_2 *are bounded open subsets of* X *with* $0 \in \Omega_1 \subset \overline{\Omega_1} \subset \Omega_2$ *and let* $\mathcal{A} : C \cap (\overline{\Omega_2} \setminus \Omega_1) \to C$ *be a completely continuous operator (continuous and compact) such that, either*

(i) $\|\mathcal{A}u\| \le \|u\|, \ \ u \in C \cap \partial\Omega_1$*, and* $\|\mathcal{A}u\| \ge \|u\|, \ \ u \in C \cap \partial\Omega_2$*, or*
(ii) $\|\mathcal{A}u\| \ge \|u\|, \ \ u \in C \cap \partial\Omega_1$*, and* $\|\mathcal{A}u\| \le \|u\|, \ \ u \in C \cap \partial\Omega_2$*.*

Then $\mathcal{A}$ *has a fixed point in* $C \cap (\overline{\Omega_2} \setminus \Omega_1)$*.*

Theorem 2. *Let* X *be a Banach space with* $\Omega \subset X$ *closed and convex. Assume* U *is a relatively open subset of* Ω *with* $0 \in U$*, and let* $S : \overline{U} \to \Omega$ *be a completely continuous operator. Then either*

(1) S *has a fixed point in* $\overline{U}$*, or*
(2) there exist $u \in \partial U$ *and* $\nu \in (0, 1)$ *such that* $u = \nu S u$*.*

3 Existence of Positive Solutions

In this section, we will give intervals for the parameter λ such that problem $(E)-(BC)$ has at least one or two positive solutions. We present now the basic assumptions that we will use in the existence results.

$(H1)$ $a_i \geq 0$ for all $i = \overline{1,p}$, $b_j \geq 0$ for all $j = \overline{1,q}$, $\sum_{i=1}^{p} a_i A^{\xi_i} \geq 1$, $\sum_{i=1}^{p} a_i \frac{1}{A^{\xi_i}} \leq 1$, $\sum_{i=1}^{q} b_i \frac{1}{A^{\eta_i}} \geq \frac{1}{A^N}$, $\sum_{i=1}^{q} b_i A^{\eta_i} \leq A^N$, $\Delta_1 > 0$ and $L > 0$.

$(H2)$ The function $f : \{1,\ldots,N-1\} \times \mathbf{R}_+ \to \mathbf{R}$ is continuous, and there exist $c_n \geq 0, n = \overline{1,N-1}$ such that $f(n,u) \geq -c_n$ for all $n = \overline{1,N-1}$, $u \in \mathbf{R}_+$, $(\mathbf{R}_+ = [0,\infty))$.

We denote by $(r_n)_{n=\overline{0,N}}$ the solution of problem (1)-(BC) with $y_n = \lambda c_n$ for all $n = \overline{1,N-1}$, namely the solution of problem

$$\begin{cases} \Delta^2 u_{n-1} - Lu_n + \lambda c_n = 0, \quad n = \overline{1,N-1}, \\ u_0 = \displaystyle\sum_{i=1}^{p} a_i u_{\xi_i}, \quad u_N = \sum_{i=1}^{q} b_i u_{\eta_i}, \end{cases} \tag{3}$$

where c_n, $n = \overline{0,N}$ are given in $(H2)$. So, by using the Green function G and Lemma 1, we have $r_n = \lambda \sum_{j=1}^{N-1} G(n,j)c_j$ for all $n = \overline{0,N}$.

We consider the difference equation

$$\Delta^2 v_{n-1} - Lv_n + \lambda(f(n,(v_n - r_n)^*) + c_n) = 0, \quad n = \overline{1,N-1}, \tag{4}$$

with the multi-point boundary conditions

$$v_0 = \sum_{i=1}^{p} a_i v_{\xi_i}, \quad v_N = \sum_{i=1}^{q} b_i v_{\eta_i}, \tag{5}$$

where $y^* = y$ if $y \geq 0$ and $y^* = 0$ if $y < 0$.

We deduce easily that the sequence $(u_n)_{n=\overline{0,N}}$ is a positive solution of problem $(E)-(BC)$ ($u_n > 0$ for all $n = \overline{0,N}$) if and only if $(v_n)_{n=\overline{0,N}}$, $v_n = u_n + r_n$, $n = \overline{0,N}$ is a solution of the boundary value problem (4)–(5) with $v_n > r_n$ for all $n = \overline{0,N}$.

By using Lemma 1, we also obtain that the sequence $(v_n)_{n=\overline{0,N}}$ is a solution of problem (4)–(5) if and only if $(v_n)_{n=\overline{0,N}}$ is a solution of the problem

$$v_n = \lambda \sum_{j=1}^{N-1} G(n,j)(f(j,(v_j - r_j)^*) + c_j), \quad n = \overline{0,N}. \tag{6}$$

We consider the Banach space $X = \mathbf{R}^{N+1} = \{v = (v_n)_{n=\overline{0,N}},\ v_n \in \mathbf{R},\ n = \overline{0,N}\}$ endowed with the maximum norm $\|v\| = \max_{n=\overline{0,N}} |v_n|$, and we define the operator $T : X \to X$, $T(v) = (T_n(v))_{n=\overline{0,N}}$, where

$$T_n(v) = \lambda \sum_{j=1}^{N-1} G(n,j)(f(j,(v_j - r_j)^*) + c_j),\quad n = \overline{0,N},\quad v = (v_n)_{n=\overline{0,N}}.$$

By $(H2)$, the operator T is completely continuous. We also define the cone

$$P = \left\{ v \in X,\quad v = (v_n)_{n=\overline{0,N}},\quad v_n \geq \frac{k(n)}{\Lambda}\|v\|,\quad \forall n = \overline{0,N} \right\}.$$

By using Lemma 3, we deduce that $T(P) \subset P$. In addition, we remark that the sequence $(v_n)_{n=\overline{0,N}}$ is a solution of problem (6) if and only if $(v_n)_{n=\overline{0,N}}$ is a fixed point of operator T. So the existence of positive solutions of problem $(E) - (BC)$ is reduced to the fixed point problem of operator T in the cone P. We denote by $k_0 = \min\{k(n),\ n = \overline{1,N-1}\}$. Because $N > 2$, we have $k_0 \in (0,1)$.

Theorem 3. *Assume that* $(H1)$, $(H2)$ *and*

$(H3)$ $\lim\limits_{u\to\infty} \min\limits_{n=\overline{1,N-1}} \dfrac{f(n,u)}{u} = \infty,$

hold. Then there exists a constant $\lambda^* > 0$ *such that for any* $\lambda \in (0,\lambda^*]$ *the problem* $(E) - (BC)$ *has at least one positive solution.*

Proof. We choose a positive number $R_1 > \frac{\Lambda^2}{k_0} \sum_{j=1}^{N-1} h(j)c_j$, and we define the set $\Omega_1 = \{v \in X,\ \|v\| < R_1\}$. We introduce the constant $\lambda^* = \min\{1, R_1(\Lambda M_0 \sum_{j=1}^{N-1} h(j))^{-1}\}$ where

$$M_0 = \max\{\max_{n=\overline{1,N-1},u\in[0,R_1]} \{f(n,u) + c_n\}, 1\}.$$

Let $\lambda \in (0,\lambda^*]$. Because $r_n \leq \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j$ for all $n = \overline{0,N}$, we deduce for any $v \in P \cap \partial\Omega_1$ that

$$[v_n - r_n]^* \leq v_n \leq \|v\| = R_1,\quad \forall n = \overline{0,N},$$

and

$$\begin{aligned} v_n - r_n &\geq \frac{k(n)}{\Lambda}\|v\| - r_n \geq \frac{k(n)}{\Lambda}\|v\| - \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j \\ &\geq \frac{k_0 R_1}{\Lambda} - \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j \geq \frac{k_0 R_1}{\Lambda} - \lambda^*\Lambda \sum_{j=1}^{N-1} h(j)c_j \\ &\geq \frac{k_0 R_1}{\Lambda} - \Lambda \sum_{j=1}^{N-1} h(j)c_j > 0,\quad \forall n = \overline{1,N-1}. \end{aligned}$$

Then for any $v \in P \cap \partial \Omega_1$, we conclude

$$
\begin{aligned}
T_n(v) &\le \lambda \Lambda \sum_{j=1}^{N-1} h(j)[f(j,(v_j - r_j)^*) + c_j] \\
&\le \lambda^* \Lambda M_0 \sum_{j=1}^{N-1} h(j) \le R_1 = \|v\|, \quad \forall n = \overline{0, N}.
\end{aligned}
$$

Hence we obtain

$$
\|T(v)\| = \max_{n=\overline{0,N}} |T_n(v)| \le \|v\|, \quad \forall v \in P \cap \partial \Omega_1. \tag{7}
$$

We consider now the constant $L_1 = 2\Lambda(\lambda k_0^2 \sum_{j=1}^{N-1} h(j))^{-1}$. By $(H3)$ we deduce that there exists a constant $M_1 > 0$ such that

$$
f(n,u) \ge L_1 u, \quad \forall n = \overline{1, N-1}, \quad u \ge M_1. \tag{8}
$$

We define $R_2 > R_1$, with $R_2 \ge \max\left\{\frac{2\Lambda M_1}{k_0}, \frac{2\Lambda^2}{k_0}\sum_{j=1}^{N-1} h(j)c_j\right\}$ and let $\Omega_2 = \{v \in X, \quad \|v\| < R_2\}$. Then for any $v \in P \cap \partial \Omega_2$, we have

$$
\begin{aligned}
v_n - r_n &= v_n - \lambda \sum_{j=1}^{N-1} G(n,j)c_j \ge v_n - \sum_{j=1}^{N-1} G(n,j)c_j \\
&\ge v_n - \Lambda \sum_{j=1}^{N-1} h(j)c_j \ge v_n - \frac{v_n \Lambda^2}{\|v\| k(n)} \sum_{j=1}^{N-1} h(j)c_j \\
&= v_n \left[1 - \frac{\Lambda^2}{R_2 k_0} \sum_{j=1}^{N-1} h(j)c_j\right] \ge \frac{v_n}{2} \ge 0, \quad \forall n = \overline{1, N-1}.
\end{aligned}
$$

So we obtain for all $n = \overline{1, N-1}$

$$
[v_n - r_n]^* = v_n - r_n \ge \frac{1}{2} v_n \ge \frac{k(n)}{2\Lambda} \|v\| \ge \frac{k_0}{2\Lambda} R_2 \ge M_1. \tag{9}
$$

Then for any $v \in P \cap \partial \Omega_2$, by (8) and (9) we deduce

$$
f(n, [v_n - r_n]^*) \ge L_1([v_n - r_n]^*) \ge \frac{L_1}{2} v_n, \quad \forall n = \overline{1, N-1}.
$$

Therefore for any $v \in P \cap \partial \Omega_2$ and $n = \overline{1, N-1}$ we conclude

$$T_n(v) \geq \lambda \sum_{j=1}^{N-1} G(n,j) L_1([v_j - r_j]^*) \geq \lambda \sum_{j=1}^{N-1} G(n,j) \frac{L_1 k_0 R_2}{2\Lambda}$$
$$\geq \frac{\lambda L_1 k_0 R_2}{2\Lambda} \sum_{j=1}^{N-1} k(n) h(j) \geq \frac{\lambda L_1 k_0^2 R_2}{2\Lambda} \sum_{j=1}^{N-1} h(j) = R_2.$$

Hence we obtain

$$\|T(v)\| \geq \|v\|, \quad \forall\, v \in P \cap \partial\Omega_2. \tag{10}$$

By (7), (10) and Theorem 1 (i), we deduce that operator T has a fixed point $v^1 = (v_n^1)_{n=\overline{0,N}} \in P$ with $R_1 \leq \|v^1\| \leq R_2$. In addition, we have

$$u_n^1 = v_n^1 - r_n = v_n^1 - \lambda \sum_{j=1}^{N-1} G(n,j) c_j \geq v_n^1 - \lambda \Lambda \sum_{j=1}^{N-1} h(j) c_j$$
$$\geq \frac{k(n)}{\Lambda} \|v^1\| - \lambda \Lambda \sum_{j=1}^{N-1} h(j) c_j \geq \frac{k_0 R_1}{\Lambda} - \Lambda \sum_{j=1}^{N-1} h(j) c_j > 0, \quad \forall n = \overline{1, N-1},$$
$$u_0^1 = v_0^1 - r_0 = \sum_{i=1}^{p} a_i u_{\xi_i}^1 > 0, \quad u_N^1 = v_N^1 - r_N = \sum_{i=1}^{q} b_i u_{\eta_i}^1 > 0.$$

Then $u^1 = (u_n^1)_{n=\overline{0,N}}$ is a positive solution of problem $(E) - (BC)$.

Theorem 4. *Assume that* $(H1)$*,* $(H2)$ *and*

$(H4)$ $\sum_{i=1}^{N-1} c_i > 0$, $\liminf_{u\to\infty} \min_{n=\overline{1,N-1}} f(n,u) > L_0$, *with* $L_0 = \frac{2\Lambda^2 \sum_{j=1}^{N-1} h(j) c_j}{k_0^2 \sum_{j=1}^{N-1} h(j)}$ *and* $\lim_{u\to\infty} \max_{n=\overline{1,N-1}} \frac{|f(n,u)|}{u} = 0$,

hold. Then there exists a constant $\lambda_* > 0$ *such that for any* $\lambda \geq \lambda_*$*, the problem* $(E) - (BC)$ *has at least one positive solution.*

Proof. By $(H4)$ we obtain that there exists $M_2 > 0$ such that $f(n,u) \geq L_0$ for all $n = \overline{1, N-1}$ and $u \geq M_2$. We define $\lambda_* = M_2 (\Lambda \sum_{j=1}^{N-1} h(j) c_j)^{-1}$. We assume that $\lambda \geq \lambda_*$. Let $R_3 = \frac{2\lambda\Lambda^2}{k_0} \sum_{j=1}^{N-1} h(j) c_j > 0$ and $\Omega_3 = \{v \in X, \ \|v\| < R_3\}$. Then for any $v \in P \cap \partial\Omega_3$, we deduce

$$v_n - r_n \geq \frac{k(n)}{\Lambda} \|v\| - \lambda \sum_{j=1}^{N-1} G(n,j) c_j \geq \frac{k(n)}{\Lambda} \|v\| - \lambda \Lambda \sum_{j=1}^{N-1} h(j) c_j$$
$$\geq \frac{k_0 R_3}{\Lambda} - \lambda \Lambda \sum_{j=1}^{N-1} h(j) c_j = \lambda \Lambda \sum_{j=1}^{N-1} h(j) c_j$$
$$\geq \lambda_* \Lambda \sum_{j=1}^{N-1} h(j) c_j = M_2 > 0, \quad \forall n = \overline{1, N-1}.$$

Therefore for any $v \in P \cap \partial \Omega_3$ we conclude

$$\begin{aligned} T_n(v) &\geq \lambda k(n) \sum_{j=1}^{N-1} h(j)(f(j,[v_j - r_j]^*) + c_j) \\ &\geq \lambda k_0 \sum_{j=1}^{N-1} h(j) L_0 = \|v\|, \quad \forall n = \overline{1, N-1}. \end{aligned}$$

So we obtain

$$\|T(v)\| \geq \|v\|, \quad \forall\, v \in P \cap \partial \Omega_3. \tag{11}$$

Next we consider the positive number $\varepsilon = (2\lambda \Lambda \sum_{j=1}^{N-1} h(j))^{-1}$. Then by $(H4)$, we deduce that there exists $M_3 > 0$ such that $|f(n,u)| \leq \varepsilon u$ for all $n = \overline{1, N-1}$ and $u \geq M_3$. So we obtain $|f(n,u)| \leq M_4 + \varepsilon u$ for all $n = \overline{1, N-1}$ and $u \geq 0$, where $M_4 = \max_{n=\overline{1,N-1}, u \in [0,M_3]} |f(n,u)|$. We define now $R_4 > R_3$ with $R_4 \geq 2\lambda \Lambda \sum_{j=1}^{N-1} h(j)(M_4 + c_j)$ and $\Omega_4 = \{v \in X, \ \ \|v\| < R_4\}$. Then for any $v \in P \cap \partial \Omega_4$, we have $v_n - r_n > M_2$ for all $n = \overline{1, N-1}$ and

$$\begin{aligned} T_n(v) &\leq \lambda \Lambda \sum_{j=1}^{N-1} h(j)[f(j,[v_j - r_j]^*) + c_j] \\ &\leq \lambda \Lambda \sum_{j=1}^{N-1} h(j)[M_4 + \varepsilon [v_j - r_j]^* + c_j] \\ &\leq \lambda \Lambda \sum_{j=1}^{N-1} h(j)(M_4 + c_j) + \lambda \varepsilon \Lambda R_4 \sum_{j=1}^{N-1} h(j) \\ &\leq \tfrac{R_4}{2} + \tfrac{R_4}{2} = R_4 = \|v\|, \quad \forall n = \overline{0, N}. \end{aligned}$$

Therefore

$$\|T(v)\| \leq \|v\|, \quad \forall\, v \in P \cap \partial \Omega_4. \tag{12}$$

By (11), (12) and Theorem 1 (ii), we deduce that operator T has a fixed point $v^1 \in P$, $v^1 = (v_n^1)_{n=\overline{0,N}}$ with $R_3 \leq \|v^1\| \leq R_4$. Besides, we conclude

$$u_n^1 = v_n^1 - r_n \ge v_n^1 - \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j \ge \frac{k(n)}{\Lambda}\|v^1\| - \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j$$
$$\ge \frac{k_0 R_3}{\Lambda} - \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j = \lambda\Lambda \sum_{j=1}^{N-1} h(j)c_j$$
$$\ge \lambda_*\Lambda \sum_{j=1}^{N-1} h(j)c_j = M_2 > 0, \ \ \forall n = \overline{1, N-1},$$
$$u_0^1 = v_0^1 - r_0 = \sum_{i=1}^{p} a_i u_{\xi_i}^1 > 0, \ \ u_N^1 = v_N^1 - r_n = \sum_{i=1}^{q} b_i u_{\eta_i}^1 > 0.$$

Then $u^1 = (u_n^1)_{n=\overline{0,N}}$ is a positive solution of problem $(E) - (BC)$.

In a similar manner used in the proof of Theorem 4, we obtain the following result.

Theorem 5. *Assume that* $(H1)$, $(H2)$ *and*

$(\widetilde{H4})$ $\sum_{i=1}^{N-1} c_i > 0$, $\lim_{u\to\infty} \min_{n=\overline{1,N-1}} f(n,u) = \infty$ *and* $\lim_{u\to\infty} \max_{n=\overline{1,N-1}} \dfrac{|f(n,u)|}{u} = 0,$

hold. Then there exists a constant $\widetilde{\lambda}_* > 0$ *such that for any* $\lambda \ge \widetilde{\lambda}_*$*, the problem* $(E) - (BC)$ *has at least one positive solution.*

Theorem 6. *Assume that* $(H1)$, $(H2)$ *and*

$(H5)$ $f(n,0) > 0$ *for all* $n = \overline{1, N-1}$,

hold. Then there exists a constant $\lambda_0 > 0$ *such that for any* $\lambda \in (0, \lambda_0]$ *the problem* $(E) - (BC)$ *has at least one positive solution.*

Proof. Let $\delta \in (0,1)$ be fixed. By using $(H2)$ and $(H5)$, there exists $R_0 \in (0,1]$ such that $f(n,u) \ge \delta f(n,0) > 0$ for all $n = \overline{1, N-1}$ and $u \in [0, R_0]$. We define

$$\overline{f}(R_0) = \max_{n=\overline{1,N-1},\, u\in[0,R_0]} \{f(n,u) + c_n\} \ge \max_{n=\overline{1,N-1}} \{\delta f(n,0) + c_n\} > 0,$$
$$\lambda_0 = R_0(2\Lambda \overline{f}(R_0) \textstyle\sum_{j=1}^{N-1} h(j))^{-1}.$$

We will show that for any $\lambda \in (0, \lambda_0]$, problem $(E) - (BC)$ has at least one positive solution. So let $\lambda \in (0, \lambda_0]$ be arbitrary, but fixed for the moment. We define the set $U = \{v \in P, \ v = (v_n)_{n=\overline{0,N}}, \ \|v\| < R_0\}$. We suppose that there exists $v \in \partial U$ $(\|v\| = R_0)$ and $\nu \in (0,1)$ such that $v = \nu T(v)$. We deduce that

$$[v_n - r_n]^* = v_n - r_n \le v_n \le R_0, \ \ if \ \ v_n - r_n \ge 0,$$
$$[v_n - r_n]^* = 0, \ \ if \ \ v_n - r_n < 0, \ \ n = \overline{0, N}.$$

Then by Lemma 2, we have

$$v_n = \nu T_n(v) \le T_n(v) \le \lambda \Lambda \sum_{j=1}^{N-1} h(j)\overline{f}(R_0) \le \lambda_0 \Lambda \overline{f}(R_0) \sum_{j=1}^{N-1} h(j) = \frac{R_0}{2}.$$

Therefore we obtain $R_0 = \|v\| \le \frac{R_0}{2}$, which is a contradiction.

Hence by Theorem 2 (with $\Omega = P$) we conclude that operator T has a fixed point $v^0 = (v_n^0)_{n=\overline{0,N}} \in \overline{U}$. That is $v^0 = T(v^0)$ or $v_n^0 = T_n(v^0)$, $n = \overline{0, N}$ and $\|v^0\| \le R_0$ with $v_n^0 \ge \frac{k(n)}{\lambda}\|v^0\|$ for all $n = \overline{0, N}$. Moreover, by Lemma 2 we deduce

$$\begin{aligned} v_n^0 = T_n(v^0) &\ge \lambda \sum_{j=1}^{N-1} G(n,j)(\delta f(j,0) + c_j) \\ &\ge \lambda k_0 \sum_{j=1}^{N-1} h(j) f(j,0) + \lambda \sum_{j=1}^{N-1} G(n,j) c_j \\ &\ge \lambda k_0 \min_{j=\overline{1,N-1}} f(j,0) \sum_{j=1}^{N-1} h(j) + r_n > r_n, \quad \forall n = \overline{1, N-1}, \end{aligned}$$

and so $u_n^0 = v_n^0 - r_n > 0$ for all $n = \overline{1, N-1}$, and $u_0^0 = v_0^0 - r_0 = \sum_{i=1}^p a_i u_{\xi_i}^0 > 0$, $u_N^0 = v_N^0 - r_N = \sum_{i=1}^q b_i u_{\eta_i}^0 > 0$. Then $u^0 = (u_n^0)_{n=\overline{0,N}}$ is a positive solution of problem $(E) - (BC)$.

Theorem 7. *Assume that* $(H1)$, $(H2)$, $(H3)$ *and* $(H5)$ *hold. Then the boundary value problem* $(E) - (BC)$ *has at least two positive solutions for* $\lambda > 0$ *sufficiently small.*

Proof. By Theorem 3 (in which we choose $R_1 > 1$) and Theorem 6, we deduce that for $0 < \lambda \le \min\{\lambda^*, \lambda_0\}$, problem $(E) - (BC)$ has at least two positive solutions $u^1 = (u_n^1)_{n=\overline{0,N}}$ and $u^0 = (u_n^0)_{n=\overline{0,N}}$ with $\|u^1 + \widetilde{r}\| > 1$ and $\|u^0 + \widetilde{r}\| \le 1$, where $\widetilde{r} = (r_n)_{n=\overline{0,N}}$.

4 Examples

Let $N = 20$, $L = 2$, $p = 2$, $q = 1$, $\xi_1 = 5$, $\xi_2 = 15$, $a_1 = 2$, $a_2 = \frac{1}{3}$, $\eta_1 = 10$, $b_1 = \frac{1}{2}$. We consider the difference equation

$$(E_0) \qquad \Delta^2 u_{n-1} - 2u_n + \lambda f(n, u_n) = 0, \quad n = \overline{1, 19},$$

with the multi-point boundary conditions

$$(BC_0) \qquad u_0 = 2u_5 + \frac{1}{3}u_{15}, \;\; u_{20} = \frac{1}{2}u_{10}.$$

We obtain $A = 2 + \sqrt{3}$, $\varDelta_1 \approx 2.73999 \times 10^{11} > 0$, $\sum_{i=1}^{p} a_i A^{\xi_i} \approx 1.26502 \times 10^8 > 1$, $\sum_{i=1}^{p} a_i \frac{1}{A^{\xi_i}} \approx 0.00276244 < 1$, $\sum_{i=1}^{q} b_i \frac{1}{A^{\eta_i}} \approx 9.53882 \times 10^{-7} > \frac{1}{A^{20}} \approx 3.63956 \times 10^{-12}$, $\sum_{i=1}^{q} b_i A^{\eta_i} \approx 262087 < A^{20} \approx 2.74758 \times 10^{11}$. Therefore assumption $(H1)$ is satisfied. In addition we have

$$\begin{aligned}
&g(n,j) = \tfrac{A}{(A^2-1)(A^{20}-A^{-20})} \begin{cases} (A^j - A^{-j})(A^{20-n} - A^{n-20}), & 1 \le j < n \le 20, \\ (A^n - A^{-n})(A^{20-j} - A^{j-20}), & 0 \le n \le j \le 19, \end{cases} \\
&G(n,j) = g(n,j) \\
&\quad + \tfrac{1}{\varDelta_1}\left[A^n\left(\tfrac{1}{2}\tfrac{1}{A^{10}} - \tfrac{1}{A^{20}}\right) + \tfrac{1}{A^n}\left(A^{20} - \tfrac{1}{2}A^{10}\right)\right]\left(2g(5,j) + \tfrac{1}{3}g(15,j)\right) \\
&\quad + \tfrac{1}{\varDelta_1}\left[A^n\left(1 - 2\tfrac{1}{A^5} - \tfrac{1}{3}\tfrac{1}{A^{15}}\right) + \tfrac{1}{A^n}\left(2A^5 + \tfrac{1}{3}A^{15} - 1\right)\right]\tfrac{1}{2}g(10,j), \\
&\qquad n = \overline{0,20}, \;\; j = \overline{1,19}, \\
&h(j) = \tfrac{A}{(A^2-1)(A^{20}-A^{-20})}(A^j - A^{-j})(A^{20-j} - A^{j-20}), \;\; j = \overline{1,19}, \\
&k(n) = \tfrac{1}{A^{19}-A^{-19}} \min\{A^n - A^{-n}, A^{20-n} - A^{n-20}\}, \;\; n = \overline{0,20}.
\end{aligned}$$

We also obtain $\varLambda \approx 3.84003043$ and $k_0 \approx 4.7053 \times 10^{-11}$.

Example 1. We consider the function

$$f(n,u) = \frac{(u+1)^{4/3}}{n(n+2)} + \ln\frac{n}{n+3}, \;\; \forall n = \overline{1,19}, \;\; u \in [0,\infty).$$

Here we have $c_n = \ln\frac{n+3}{n} \ge 0$ for all $n = \overline{1,19}$ and then we obtain $f(n,u) \ge -c_n$ for all $n = \overline{1,19}$ and $u \in [0,\infty)$. Because $\lim_{u\to\infty} f(n,u)/u = \infty$, then the assumptions $(H2)$ and $(H3)$ are satisfied. We choose $R_1 = 6.53502 \times 10^{11}$ which satisfies the condition from the beginning of the proof of Theorem 3. Then $M_0 \approx 1.89034 \times 10^{15}$ and $\lambda^* \approx 3.09645 \times 10^6$. By Theorem 3 we deduce that $(E_0) - (BC_0)$ has at least one positive solution for any $\lambda \in (0, \lambda^*]$.

Example 2. We consider the function

$$f(n,u) = (u-1)(u-2), \;\; n = \overline{1,19}, \;\; u \in [0,\infty).$$

Then there exists $\widetilde{M}_0 > 0$ ($\widetilde{M}_0 = \frac{1}{4}$) such that $f(n,u) + \widetilde{M}_0 \ge 0$ for all $n = \overline{1,19}$ and $u \ge 0$, ($c_n = \widetilde{M}_0 = \frac{1}{4}$ for all $n = \overline{1,19}$) and $f(n,0) = 2 > 0$ for all $n = \overline{1,19}$. So assumptions $(H2)$ and $(H5)$ are satisfied. Let $\delta = \frac{3}{8} < 1$ and $R_0 = \frac{1}{2}$. Then $f(n,u) \ge \delta f(n,0)$ for all $n = \overline{1,19}$ and $u \in [0, \frac{1}{2}]$. Besides $\overline{f}(R_0) = \frac{9}{4}$ and therefore we obtain $\lambda_0 \approx 9.95213 \times 10^8$. By Theorem 6, for any $\lambda \in (0, \lambda_0]$, we deduce that problem $(E_0) - (BC_0)$ has at least one positive solution.

References

1. Afrouzi, G.A., Hadjian, A.: Existence and multiplicity of solutions for a discrete nonlinear boundary value problem. Electr. J. Differ. Equ. **2014**(35), 1–13 (2014)
2. Agarwal, R.P.: Difference equations and inequalities: theory, methods, and applications. Second edn. Monographs and Textbooks in Pure and Applied Mathematics, vol. 228. Marcel Dekker, Inc., New York (2000)
3. Agarwal, R.P., Luca, R.: Positive solutions for a system of second-order discrete boundary value problems. Adv. Differ. Equ. **2018**(470), 1–17 (2018)
4. Agarwal, R.P., Meehan, M., O'Regan, D.: Fixed Point Theory and Applications. Cambridge University Press, Cambridge (2001)
5. Anderson, D.R.: Solutions to second-order three-point problems on time scales. J. Diff. Equ. Appl. **8**(8), 673–688 (2002)
6. Atici, F., Peterson, A.C.: Inequality for a $2n$th order difference equation. PanAmer. Math. J. **6**, 41–49 (1996)
7. Avery, R.: Three positive solutions of a discrete second order conjugate problem. PanAmer. Math. J. **8**, 79–96 (1998)
8. Bai, D., Henderson, J., Zeng, Y.: Positive solutions of discrete Neumann boundary value problems with sign-changing nonlinearities. Bound. Value Probl. **2015**(231), 1–9 (2015)
9. Cheung, W.S., Ren, J.: Positive solution for discrete three-point boundary value problems. Aust. J. Math. Anal. Appl. **1**(2), 1–7 (2004). Article 9
10. Graef, J.R., Kong, L., Wang, M.: Multiple solutions to a periodic boundary value problem for a nonlinear discrete fourth order equation. Adv. Dyn. Syst. Appl. **8**(2), 203–215 (2013)
11. Graef, J.R., Kong, L., Wang, M.: Existence of multiple solutions to a discrete fourth order periodic boundary value problem. Discrete Cont. Dyn. Syst. **2013**(Supplement), 291–299 (2013)
12. Guo, D., Lakshmikantham, V.: Nonlinear Problems in Abstract Cones. Academic Press, New York (1988)
13. Henderson, J., Luca, R.: Positive solutions for a system of difference equations with coupled multi-point boundary conditions. J. Diff. Equ. Appl. **22**(2), 188–216 (2016)
14. Henderson, J., Luca, R.: Boundary Value Problems for Systems of Differential, Difference and Fractional Equations. Positive Solutions. ElsevierElsevier, Amsterdam (2016)
15. Henderson, J., Luca, R.: Existence and multiplicity of positive solutions for a system of difference equations with coupled boundary conditions. J. Appl. Anal. Comput. **7**(1), 134–146 (2017)
16. Henderson, J., Luca, R., Tudorache, A.: Multiple positive solutions for a multi-point discrete boundary value problem. Commun. Fac. Sci. Univ. Ankara Ser. A1 Math. Stat. **63**(2), 59–70 (2014)
17. Henderson, J., Luca, R., Tudorache, A.: Existence and nonexistence of positive solutions to a discrete boundary value problem. Carpathian J. Math. **33**(2), 181–190 (2017)
18. Kelley, W.G., Peterson, A.C.: Difference Equations. An Introduction with Applications, 2nd edn. Academic Press, San Diego (2001)
19. Lakshmikantham, V., Trigiante, D.: Theory of Difference Equations Numerical Methods and Applications. Mathematics in Science and Engineering, vol. 181. Academic Press Inc, Boston (1988)
20. Luca, R.: Positive solutions for a semipositone nonlocal discrete boundary value problem. Appl. Math. Lett. **92**, 54–61 (2019)
21. Luca, R.: Existence of positive solutions for a semipositone discrete boundary value problem. Nonlinear Anal. Model. Control **24**(4), 658–678 (2019)
22. Rodriguez, J.: Nonlinear discrete Sturm-Liouville problems. J. Math. Anal. Appl. **308**(1), 380–391 (2005)
23. Wang, D.B., Guan, W.: Three positive solutions of boundary value problems for p-Laplacian difference equations. Comp. Math. Appl. **55**(9), 1943–1949 (2008)
24. Wang, L., Chen, X.: Positive solutions for discrete boundary value problems to one-dimensional p-Laplacian with delay. J. Appl. Math. **2013**, 1–8 (2013). Article ID 157043

Dynamics of a Certain Nonlinearly Perturbed Heat Equation

Carlos Ramos, Ana Isabel Santos, and Sandra Vinagre

Abstract We consider a system described by the linear heat equation, with appropriate boundary conditions in order to model the temperature on a wire with adiabatic endpoints, which is perturbed nonlinearly by a family of quadratic maps. The time instants of the perturbation are determined by an additional dynamical system, seen here as part of the external interacting system. We study the complex behaviour of the system, namely the dependence on initial conditions.

1 Introduction

We study an infinite dimension dynamical system whose irregular behavior is mainly determined by an iterated map of the interval. Previous works with similar perspective and which were our motivation can be found in [6–9] and [10]. The configuration of the system considered here is itself an interval, I, and the state of the system is determined by a differentiable function on I. For the time evolution we consider two distinct phases: a regular continuous phase determined by the diffusion equation (heat equation), characterized by a diffusing coefficient λ; a perturbation regime, with discrete time, which is determined by the action of an operator, T_{f_μ}, induced by an interval map f_μ, characterized by a parameter μ. The dissipative nature of the diffusion equation progressively eliminates irregular aspects of the state function with the time flow. Therefore, we expect rapid convergent behavior to a stable state. However, the external perturbation, in this case modeled by T_{f_μ}, if it arises from a chaotic interval map f_μ, introduces irregularities and complexity for the possible observed states, in particular the number of critical points of the state functions grows

C. Ramos (✉) · A. I. Santos · S. Vinagre
Department of Mathematics, ECT, CIMA, University of Évora,
Rua Romão Ramalho 59, 7000-671 Évora, Portugal
e-mail: ccr@uevora.pt

A. I. Santos
e-mail: aims@uevora.pt

S. Vinagre
e-mail: smv@uevora.pt

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_47

and evolves in a chaotic manner. The interplay between the dissipative nature of the heat equation and the expansive nature of the interval map gives us a very complex behavior, taking into account the variation of the control parameters λ and μ. We have previously addressed this problem, see [2, 4] and [5], for the fully developed chaotic regime, with μ so that the topological entropy of f_μ is positive. In this case, we observe a that a transient phase occurs, in which the number of critical points of the state function grows exponentially with time, for almost every initial condition. After these transient phase the state functions oscillate in a chaotic manner, nevertheless with a limited number of critical points. This number of critical points vary around a certain average value. This average number of critical points and also its standard deviation are mainly dependent on the parameter λ and not on μ, see [4]. Moreover, this stationary average number of critical points, in the stationary phase, does not depend on the initial condition chosen, only the time needed to attain the stationary phase.

Here, we deal with the zero topological entropy situation. In this case, we observe that the number of critical points in the stationary phase depends strongly on the initial condition. Moreover, although the topological entropy of the interval map is zero, which would lead us to expect simple behavior—in fact all the initial conditions lead to periodic state functions—nevertheless, there are an infinite number of attractive periodic functions, depending on the initial condition. Therefore, this case deserves further study, which is initiated here. We start with certain preliminary material.

Consider the class of differentiable functions

$$\mathscr{A} = \left\{\varphi \in C^1([0,1]) : \varphi'(0) = \varphi'(1) = 0, |cp\,(\varphi)| < \infty\right\},$$

where $|cp\,(\varphi)|$ denotes the number of critical points of φ. That is, a function belongs to the class $\mathscr{A}$ if it is differentiable, its derivatives at the endpoints are 0 and its number of critical points are finite.

Following [2], the interval maps we consider here, modeling the perturbation, belongs to the well studied quadratic family defined by $f_\mu\,(x) = 1 - \mu x^2$, with $\mu \in (0, 2]$. There is a maximal invariant interval, $[-1, 1]$, where the relevant dynamics occurs, that is, the iterates $f_\mu^k\,(x_0) := f_\mu\left(\ldots f_\mu\,(x_0)\right)$ (k times) of initial points x_0 in $[-1, 1]$ will belong to $[-1, 1]$, for every k. For initial points x_0 outside $[-1, 1]$, the iterates $f_\mu^k\,(x_0)$, $k \in \mathbb{N}$ diverge to infinity.

The one parameter family of operators T_{f_μ}, induced by f_μ on $\mathscr{A}$, which determines the perturbation regime, is defined by

$$\begin{aligned} T_{f_\mu} : \mathscr{A} &\to \quad \mathscr{A} \\ \varphi &\mapsto f_\mu \circ \varphi. \end{aligned}$$

For each parameter μ, the operator T_{f_μ} is well defined since $\left(f_\mu \circ \varphi\right)'(0) = \left(f_\mu \circ \varphi\right)'(1) = 0$, therefore, we consider the discrete dynamical system $(\mathscr{A}, T_{f_\mu})$. Since points outside the interval $[-1, 1]$ goes to infinity under iteration of f_μ, and every iterate of initial points belonging to the interval $[-1, 1]$ belongs to $[-1, 1]$, the iterates of initial functions ϕ_0 whose image is contained in $[-1, 1]$ will maintain its image contained in $[-1, 1]$. If the image of ϕ_0 is not contained in $[-1, 1]$, then the iterates of ϕ_0 under the operator T_{f_μ} will explode.

2 Nonlinear Perturbed Heat Equation

We consider the unit interval representing an ideal wire. The temperature function at each point $x \in [0, 1]$ and each time instant $t \in \mathbb{R}_0^+$ is denoted by $\psi(x, t)$. We consider also that the wire is such that the time evolution of the temperature function is described by the linear heat equation

$$\frac{\partial \psi}{\partial t} = \lambda \frac{\partial^2 \psi}{\partial x^2}, \tag{1}$$

where λ is a constant, the *diffusion coefficient*. If there is no heat exchange in the endpoints $x = 0$ and $x = 1$, we have adiabatic boundary conditions

$$\frac{\partial \psi}{\partial x}(0, t) = \frac{\partial \psi}{\partial x}(1, t) = 0. \tag{2}$$

The initial condition $\psi(x, 0) = \phi_0(x)$ is chosen from the class $\mathscr{A}$, introduced in the previous section. The solution can be written as follows

$$\psi(x, t) = \sum_{n=0}^{\infty} c_n e^{-n^2\pi^2\lambda t} \cos(n\pi x), \tag{3}$$

where the coefficients c_n are determined by the initial condition written as a cosine Fourier series

$$\phi_0(x) = \sum_{n=0}^{\infty} c_n \cos(n\pi x). \tag{4}$$

From the explicit solution (3), we see that for every fixed $t_* \in \mathbb{R}_0^+$, we have $\psi(x, t_*) \in \mathscr{A}$.

Suppose the system is perturbed in time instants $t_1, t_2, \ldots$ through a certain nonlinear process. Being the temperature distribution along the wire initially given by the function $\psi_0(x, t)$, for $t_0 < t < t_1$, after the perturbation the temperature function is $\psi_1(x, t)$, for $t > t_1$. We have continuous time evolution for $t \in]t_j, t_{j+1}[$ and discrete time evolution for $t = t_j$. We assume that the perturbation is characterized by a nonlinear map f so that $\psi_{j+1}(x, t_{j+1}) = f(\psi_j(x, t_{j+1}))$, with $\psi_1(x, t_1) = f(\psi_0(x, t_1))$. If the time instants are $t_k = k \in \mathbb{N}$, the time evolution of the system is described by the sequence of functions

$$\{\psi_0, \psi_1, \psi_2, \ldots, \psi_k, \ldots\}, \tag{5}$$

each function ψ_k satisfying the heat equation for $x \in [0, 1]$, $t \in [k, k + 1[$, $k \in \mathbb{N}_0$, with initial conditions determined by

$$\psi_{k+1}(x, k + 1) = f\,(\psi_k(x, k + 1))\,, \text{ for } k \in \mathbb{N}_0,$$

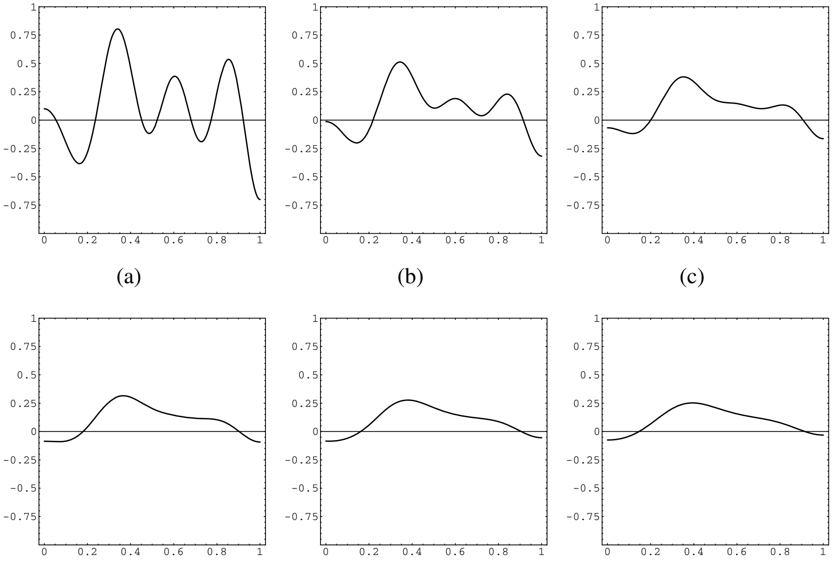
and $\psi_0(x, 0) = \phi_0\,(x)$ a given function from $\mathscr{A}$.

Next, in a first example, we show snapshots of the evolution of the system in a linear continuous regime, still without perturbation, and, in a second example, we show what occurs to the system when we introduce a perturbation at certain time instants.

Example 1. Let us consider $\lambda = 0.005$ and

$$\begin{aligned}\psi_0\,(x, 0) = \phi_0\,(x) &= 0.1 - 0.2\cos(2\pi x) - 0.1\cos(3\pi x) - 0.1\cos(4\pi x)\\ &\quad +0.2\cos(5\pi x) + 0.1\cos(6\pi x) + .3\cos(7\pi x) - 0.2\cos(8\pi x).\end{aligned}$$

In Fig. 1, we show the evolution of the system in a linear continuous regime for the initial condition $\psi\,(x, 0)$.



(a) (b) (c)

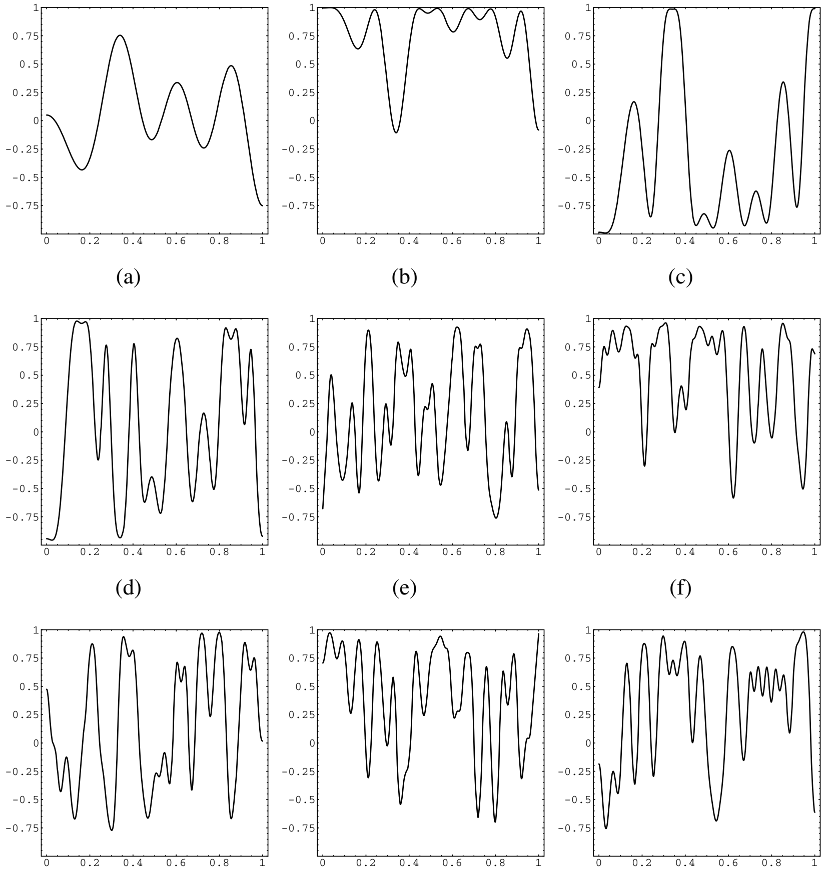
(d) (e) (f)

Fig. 1 Graphs of **a** $\psi\,(x, 0)$, **b** $\psi\,(x, 0.4)$, **c** $\psi\,(x, 0.8)$, **d** $\psi\,(x, 1.2)$, **e** $\psi\,(x, 1.6)$ and **f** $\psi\,(x, 2)$, with $\lambda = 0.005$ and $\psi\,(x, 0) = 0.1 - 0.2\cos(2\pi x) - 0.1\cos(3\pi x) - 0.1\cos(4\pi x) + 0.2\cos(5\pi x) + 0.1\cos(6\pi x) + .3\cos(7\pi x) - 0.2\cos(8\pi x)$

Example 2. Consider $f_\mu(x) = 1 - \mu x^2$, with $\mu = 2$, $\lambda = 0.00005$ and

$$\psi_0(x,0) = \phi_0(x) = 0.1 - 0.2\cos(2\pi x) - 0.1\cos(3\pi x) - 0.1\cos(4\pi x) \\ +0.2\cos(5\pi x) + 0.1\cos(6\pi x) + .3\cos(7\pi x) - 0.2\cos(8\pi x).$$

In the Fig. 2, we show the evolution of the system described by the heat equation, which is perturbed in time instants $t = 1, 2, 3, 4, 90, 91, 92, 93, 94$.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 2 Graphs of **a** $\psi_0(x,1)$, **b** $\psi_1(x,2)$, **c** $\psi_2(x,3)$, **d** $\psi_3(x,4)$, **e** $\psi_{90}(x,91)$, **f** $\psi_{91}(x,92)$, **g** $\psi_{92}(x,93)$, **h** $\psi_{93}(x,94)$ and **i** $\psi_{94}(x,95)$, with $\lambda = 0.00005$, $f_\mu(x) = 1 - \mu x^2$, $\mu = 2$ and $\psi(x,0) = 0.1 - 0.2\cos(2\pi x) - 0.1\cos(3\pi x) - 0.1\cos(4\pi x) + 0.2\cos(5\pi x) + 0.1\cos(6\pi x) + .3\cos(7\pi x) - 0.2\cos(8\pi x)$

The discrete dynamical system used in this work is the following. We consider the state space $\mathscr{A}$, the operator T_{f_μ} and an operator $U_{\lambda,\varepsilon} : \mathscr{A} \to \mathscr{A}$, which gives the time evolution under the unperturbed regime, with diffusion coefficient λ. The operator $U_{\lambda,\varepsilon}$ is defined implicitly by

$$U_{\lambda,\varepsilon}\psi\,(x,t) := \psi\,(x,t+\varepsilon)\,.$$

Let us consider the operator $V_{\mu,\lambda,\varepsilon} : \mathscr{A} \to \mathscr{A}$ defined by

$$V_{\mu,\lambda,\varepsilon} := T_{f_\mu} \circ U_{\lambda,\varepsilon}.$$

If the system is perturbed in natural time instants, with fixed increments, it is sufficient to consider a natural value for ε. In fact, if the time increment is $a \notin \mathbb{N}$, we can rescale through the parameter λ. Therefore, we set $\varepsilon = 1$ and we define $V_{\mu,\lambda} \equiv V_{\mu,\lambda,1}$. Our discrete dynamical system is, then, defined by the pair $\left(\mathscr{A}, V_{\mu,\lambda}\right)$. When we iterate a function $\phi_0\,(x)$ in $\mathscr{A}$, under $V_{\mu,\lambda}$, the obtained iterates $\phi_k\,(x) = V^k_{\mu,\lambda}\,(\phi_0\,(x))$ will correspond to the solution given by the sequence of functions (5) in the time instants $\phi_k\,(x) = \psi_k\,(x,k)$. If, for some reason, we need to obtain the temperature function at a non integer time instant t' we simply use the solution presented in (3) with initial condition given by $\psi\,(x,0) = V^k_{\mu,\lambda}\,(\phi_0\,(x))$, where $k = \left[t'\right]$ is the integer part of t'. Then, we evaluate the function for the time instant $t' - k$, that is, $\psi\,\left(x, t' - k\right)$.

3 The Evolution of Critical Points of the Iterates $\phi_k = V^k_{\mu,\lambda}\,(\phi_0)$

As we referred above, in our previous works was established that, in certain conditions—namely positive topological entropy of f_μ—the iterates, under $V_{\mu,\lambda}$, have an exponential grow of number of critical points up to a certain level, see [2]. After attaining a certain number of critical points, which depends on the parameters, this number oscillates and becomes limited. For the dynamical system $\left(\mathscr{A}, V_{\mu,\lambda}\right)$, the number of critical points does not grow exponentially. Indeed, it attains a balance between the creation of new critical points, due to the interval map effect, and the destruction of critical points, due to the dissipative effect of the heat equation.

The topological entropy of an interval map g is an important measure for the characterization of the complex behaviour of the map under iteration (see [3]). Roughly speaking as greater topological entropy more complex is the dynamical behaviour. For the infinite dimensional system $(\mathscr{A}, T_g)$, the topological entropy of g measures the growth rate of the number of critical points for the functions in $\mathscr{A}$ (see [1]). That is, if the topological entropy is zero, then the growth rate is polynomial or there is no growth at all, for almost every initial functions. For positive topological entropy the iterates will have an increasingly number of critical points, growing exponentially under iteration.

In what follows we consider a set of representative initial conditions, whose graphs are in the Fig. 7:

$$\begin{aligned}\psi_0(x) &= 0.2 + 0.1\cos(\pi x) - 0.2\cos(2\pi x) + 0.1\cos(3\pi x)\\ &\quad + 0.1\cos(4\pi x) - 0.1\cos(5\pi x) + 0.2\cos(6\pi x),\\ \widetilde{\psi}_0(x) &= 0.2 + 0.1\cos(\pi x) - 0.2\cos(2\pi x) + 0.1\cos(3\pi x),\\ \varphi_0(x) &= 0.2 + 0.1\cos(\pi x) - 0.2\cos(2\pi x) + 0.1\cos(3\pi x)\\ &\quad + 0.1\cos(4\pi x) - 0.1\cos(5\pi x) + 0.2\cos(6\pi x)\\ &\quad + 0.1\cos(7\pi x) + 0.2\cos(8\pi x) + 0.2\cos(9\pi x),\\ \widetilde{\varphi}_0(x) &= 0.03\cos(2\pi x) + 0.2\cos(3\pi x) - 0.2\cos(5\pi x)\\ &\quad + 0.3\cos(6\pi x) - 0.3\cos(29\pi x),\\ \phi_0(x) &= 0.07\cos(2\pi x) + 0.07\cos(3\pi x) - 0.07\cos(5\pi x)\\ &\quad + 0.07\cos(6\pi x) - 0.07\cos(29\pi x),\\ \widetilde{\phi}_0(x) &= 0.03\cos(2\pi x) + 0.05\cos(3\pi x) - 0.07\cos(4\pi x)\\ &\quad + 0.07\cos(5\pi x) + 0.03\cos(7\pi x),\end{aligned}$$

$$\rho_0(x) = 0.45 - 0.45\cos(2\pi x),$$

$$\widetilde{\rho}_0(x) = -0.75\cos(4\pi x),$$

$$\alpha_0(x) = 0.3 + 0.6\cos(2\pi x) - 0.45\cos(4\pi x),$$

$$\widetilde{\alpha}_0(x) = -0.1 - 0.55\cos(2\pi x) - 0.2\cos(4\pi x) + 0.5\cos(6\pi x),$$

$$\beta_0(x) = -0.1 - 0.55\cos(2\pi x) - 0.15\cos(4\pi x) - 0.15\cos(6\pi x) + 0.5\cos(8\pi x),$$

$$\widetilde{\beta}_0(x) = -0.75\cos(6\pi x),$$

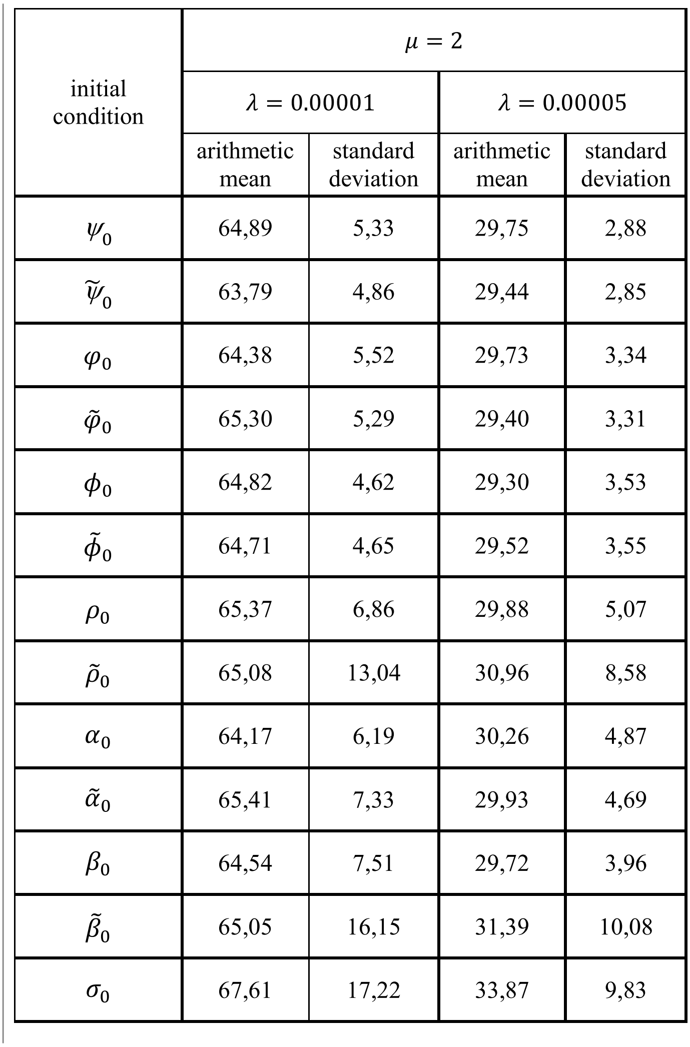
and

$$\sigma_0(x) = -0.75\cos(8\pi x).$$

3.1 Positive Topological Entropy of f_μ

The average number of critical points of the iterates depends on the diffusion coefficient λ, when we decrease the diffusion coefficient λ, the average number of critical points, at which the temperature function stabilizes, increases, as we can see in the Table 1.

Table 1 The arithmetic mean and the standard deviation for thirteen different initial conditions, ψ_0, $\widetilde{\psi}_0$, φ_0, $\widetilde{\varphi}_0$, ϕ_0, $\widetilde{\phi}_0$, ρ_0, $\widetilde{\rho}_0$, α_0, $\widetilde{\alpha}_0$, β_0, $\widetilde{\beta}_0$, σ_0, of the number of critical points. The values are presented for $\mu = 2$ and for two different values of the diffusion coefficient λ (λ = 0.00005 and λ = 0.00001)



initial condition	$\mu = 2$			
	$\lambda = 0.00001$		$\lambda = 0.00005$	
	arithmetic mean	standard deviation	arithmetic mean	standard deviation
ψ_0	64,89	5,33	29,75	2,88
$\widetilde{\psi}_0$	63,79	4,86	29,44	2,85
φ_0	64,38	5,52	29,73	3,34
$\tilde{\varphi}_0$	65,30	5,29	29,40	3,31
ϕ_0	64,82	4,62	29,30	3,53
$\tilde{\phi}_0$	64,71	4,65	29,52	3,55
ρ_0	65,37	6,86	29,88	5,07
$\tilde{\rho}_0$	65,08	13,04	30,96	8,58
α_0	64,17	6,19	30,26	4,87
$\tilde{\alpha}_0$	65,41	7,33	29,93	4,69
β_0	64,54	7,51	29,72	3,96
$\tilde{\beta}_0$	65,05	16,15	31,39	10,08
σ_0	67,61	17,22	33,87	9,83

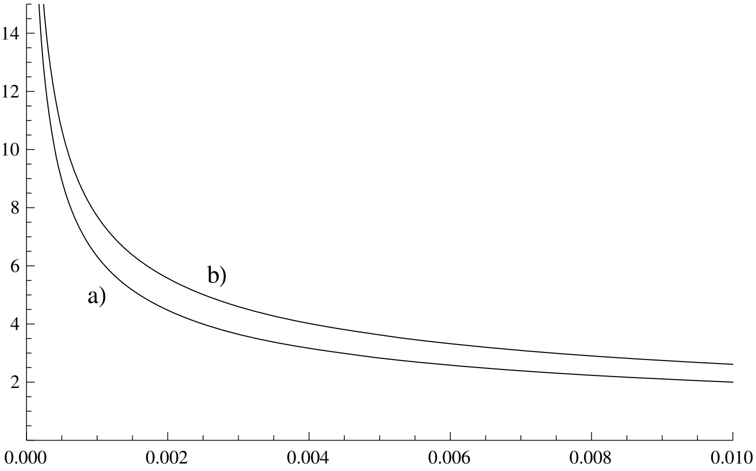


Fig. 3 Graph of $\eta(\lambda) = C_0\lambda^{C_1}$, with **a** $C_0 = 0.2$ and $C_1 = -0.5$ and **b** $C_0 = 0.3$ and $C_1 = -0.47$

Using the values obtained numerically for the distribution of the critical points, we determined, in [4], an empirical relation which characterizes this phenomena, and we can enunciate the following result.

Numerical Result 1. *When μ is such that $h_t\left(f_\mu\right) > 0$, the average number of critical points, $\eta(\lambda)$, does not depend on μ, only on λ. Moreover, if $\lambda \in [0.0000075, 0.01]$, then we have approximately the rule*

$$\eta(\lambda) = C_0\lambda^{C_1}, \tag{6}$$

with $C_0 = 0.25 \pm 0.05$ and $C_1 = -0.485 \pm 0.015$ (see Fig. 3).

3.2 Topological Entropy of f_μ Equal to Zero

For the cases in which the quadratic map f_μ has topological entropy equal to zero, that is, the cases the parameter μ is such that $h_t\left(f_\mu\right) = 0$, it is not possible to present a formula similar to the one that was presented in the result (6). For zero topological entropy the number and the evolution of critical points depends strongly on the initial conditions, regardless of the value of λ. Moreover, the standard deviation is very high although the growth of the critical points is low, as we can see in the Table 2. However, we observed that for fixed initial conditions the asymptotic behavior follows a pattern similar to the previously identified in the result (6). In this case, what depends on the initial condition, is not the existence of exponential decay on the parameter λ, are the constants of this decay, as we enunciate in the result (7).

Table 2 The arithmetic mean and the standard deviation for thirteen different initial conditions, ψ_0, $\widetilde{\psi}_0$, φ_0, $\widetilde{\varphi}_0$, ϕ_0, $\widetilde{\phi}_0$, ρ_0, $\widetilde{\rho}_0$, α_0, $\widetilde{\alpha}_0$, β_0, $\widetilde{\beta}_0$, σ_0, of the number of critical points. The values are presented for $\mu = 1.3815\ldots$ and for five different values of the diffusion coefficient λ ($\lambda = 0.0001$, $\lambda = 0.00005$, $\lambda = 0.00003$, $\lambda = 0.00001$ and $\lambda = 0.0000075$)

initial condition	$\mu = 1.3815$ ...									
	$\lambda = 0.0000075$		$\lambda = 0.00001$		$\lambda = 0.00003$		$\lambda = 0.00005$		$\lambda = 0.0001$	
	arithmetic mean	standard deviation	arithmetic mean	standard deviation	arithmetic mean	standard deviation	arithmetic mean	standard deviation	arithmetic mean	standard deviation
ψ_0	31,25	4,68	28,23	5,50	18,28	3,35	16,55	2,28	12,86	1,96
$\widetilde{\psi}_0$	30,93	7,00	29,76	9,71	17,45	4,95	13,55	3,25	6,77	1,46
φ_0	40,23	5,16	35,14	3,45	21,20	4,10	15,11	3,09	13,40	1,29
$\tilde{\varphi}_0$	46,09	6,51	41,09	6,52	20,90	4,33	15,37	5,95	10,70	2,45
ϕ_0	22,76	4,22	22,61	7,51	18,16	8,89	14,41	6,17	9,72	2,62
$\tilde{\phi}_0$	44,01	3,93	35,58	4,06	11,58	2,94	7,75	2,46	7,22	3,18
ρ_0	39,11	8,26	33,17	6,13	18,59	2,33	15,02	2,19	10,07	2,29
$\tilde{\rho}_0$	46,98	7,06	43,08	9,95	32,46	2,34	23,95	6,09	19,34	5,24
α_0	42,15	3,65	36,96	4,72	27,67	3,08	23,10	3,50	17,69	2,93
$\tilde{\alpha}_0$	48,39	6,05	42,54	4,53	30,58	3,21	22,84	3,13	18,63	2,49
β_0	47,44	4,87	43,37	4,79	27,08	4,08	22,49	3,44	15,18	1,84
$\tilde{\beta}_0$	57,50	11,08	56,00	9,75	30,57	6,46	25,24	8,35	16,18	5,40
σ_0	62,86	7,49	54,77	9,52	33,00	11,14	24,76	8,00	-	-

Table 3 The values for C_0 and C_1, em (7), for the initial conditions ψ_0, $\widetilde{\psi}_0$, φ_0, $\widetilde{\varphi}_0$, ϕ_0, $\widetilde{\phi}_0$, ρ_0, $\widetilde{\rho}_0$, α_0, $\widetilde{\alpha}_0$, β_0, $\widetilde{\beta}_0$ and σ_0

initial condition	C_0	C_1
ψ_0	0.6368±0.0295	-0.3266±0.0031
$\widetilde{\psi}_0$	0.0416±0.0055	-0.5722±0.0007
φ_0	0.1552±0.0300	-0.4714±0.0107
$\tilde{\varphi}_0$	0.0337±0.0138	-0.6222±0.0307
ϕ_0	0.6219±0.1847	-0.3163±0.0113
$\tilde{\phi}_0$	0.0053±0.0035	-0.7780±0.0615
ρ_0	0.0791±0.0079	-0.5253±0.0012
$\tilde{\rho}_0$	0.7169±0.1649	-0.3612±0.0132
α_0	0.8717±0.1429	-0.3309±0.0098
$\tilde{\alpha}_0$	0.5912±0.0891	-03737±0.0083
β_0	0.2891±0.0433	-0.4351±0.0080
$\tilde{\beta}_0$	0.1785±0.0797	-0.5056±0.0346
σ_0	0.2010±0.1257	-0.5079±0.0566

Numerical Result 2. *Let $\psi_0 \in A$ and μ such that $h_t\left(f_\mu\right) = 0$. If $\lambda \in [0.0000075, 0.0001]$, then the average number of critical points, $\eta_{\mu,\psi_0}(\lambda)$, is given approximately by the rule*

$$\eta_{\mu,\psi_0}(\lambda) = C_0(\psi_0)\,\lambda^{C_1(\psi_0)}, \tag{7}$$

where $C_0(\psi_0)$ and $C_1(\psi_0)$ depend on the initial condition ψ_0 and μ .

In the Table 3, we present the values for C_0 and C_1, em (7), for the initial conditions ψ_0, $\widetilde{\psi}_0$, φ_0, $\widetilde{\varphi}_0$, ϕ_0, $\widetilde{\phi}_0$, ρ_0, $\widetilde{\rho}_0$, α_0, $\widetilde{\alpha}_0$, β_0, $\widetilde{\beta}_0$ and σ_0, and in the following example

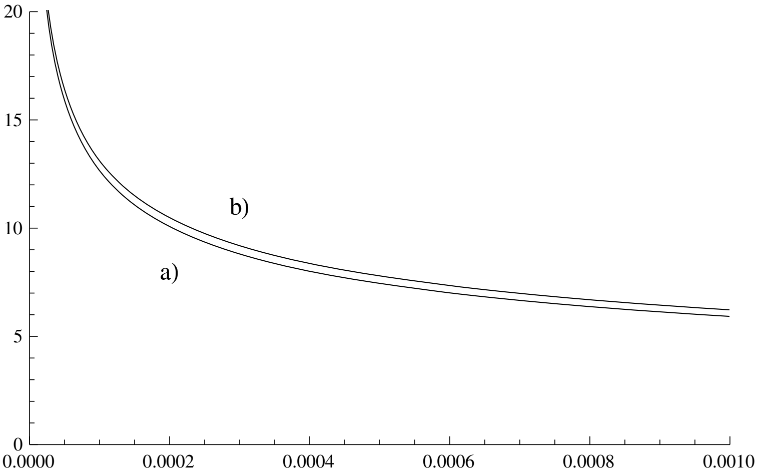


Fig. 4 Graph of $\eta_{\mu,\psi_0}(\lambda) = C_0(\psi_0)\,\lambda^{C_1(\psi_0)}$, with **a** $C_0(\psi_0) = 0.6073$ and $C_1(\psi_0) = -0.3297$ and **b** $C_0(\psi_0) = 0.6663$ and $C_1(\psi_0) = -0.3235$

we present the graph of the average number of the critical points for three of these initial conditions.

Example 3. Consider $f_\mu(x) = 1 - \mu x^2$, with $\mu = 1.3815\ldots$,

$$\begin{aligned}\psi_0(x,0) &= 0.2 + 0.1\cos(\pi x) - 0.2\cos(2\pi x) + 0.1\cos(3\pi x)\\ &\quad + 0.1\cos(4\pi x) - 0.1\cos(5\pi x) + 0.2\cos(6\pi x),\end{aligned}$$

$$\widetilde{\psi}_0(x,0) = 0.2 + 0.1\cos(\pi x) - 0.2\cos(2\pi x) + 0.1\cos(3\pi x)$$

and

$$\rho_0(x,0) = 0.45 - 0.45\cos(2\pi x).$$

For these initial conditions we have $C_0(\psi_0) = 0.6368 \pm 0.0295$ and $C_1(\psi_0) = -0.3266 \pm 0.0031$, $C_0(\varphi_0) = 0.0416 \pm 0.0055$ and $C_1(\varphi_0) = -0.5722 \pm 0.0007$, and $C_0(\rho_0) = 0.0791 \pm 0.0079$ and $C_1(\rho_0) = -0.5253 \pm 0.0012$. In Figs. 4, 5 and 6, we present the graphs of $\eta_{\mu,\psi_0}(\lambda)$.

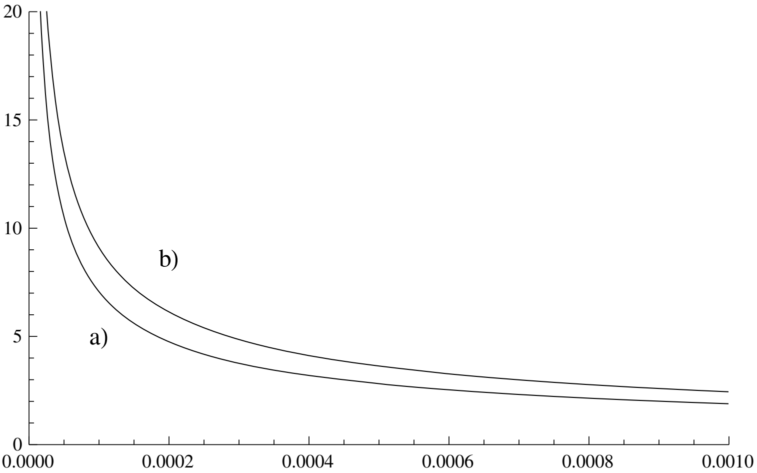


Fig. 5 Graph of $\eta_{\mu,\varphi_0}(\lambda) = C_0(\varphi_0)\,\lambda^{C_1(\varphi_0)}$, with **a** $C_0(\varphi_0) = 0.0361$ and $C_1(\varphi_0) = -0.5729$ and **b** $C_0(\varphi_0) = 0.0471$ and $C_1(\varphi_0) = -0.5715$

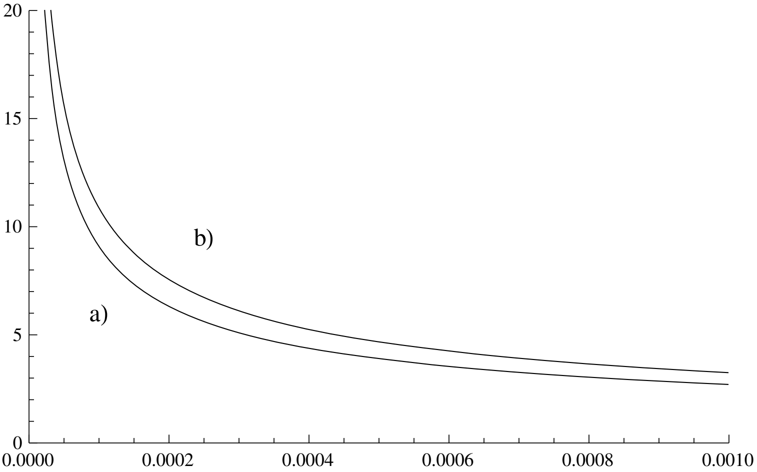


Fig. 6 Graph of $\eta_{\mu,\rho_0}(\lambda) = C_0(\rho_0)\,\lambda^{C_1(\rho_0)}$, with **a** $C_0(\rho_0) = 0.0712$ and $C_1(\rho_0) = -0.5265$ and **b** $C_0(\rho_0) = 0.087$ and $C_1(\rho_0) = -0.5241$

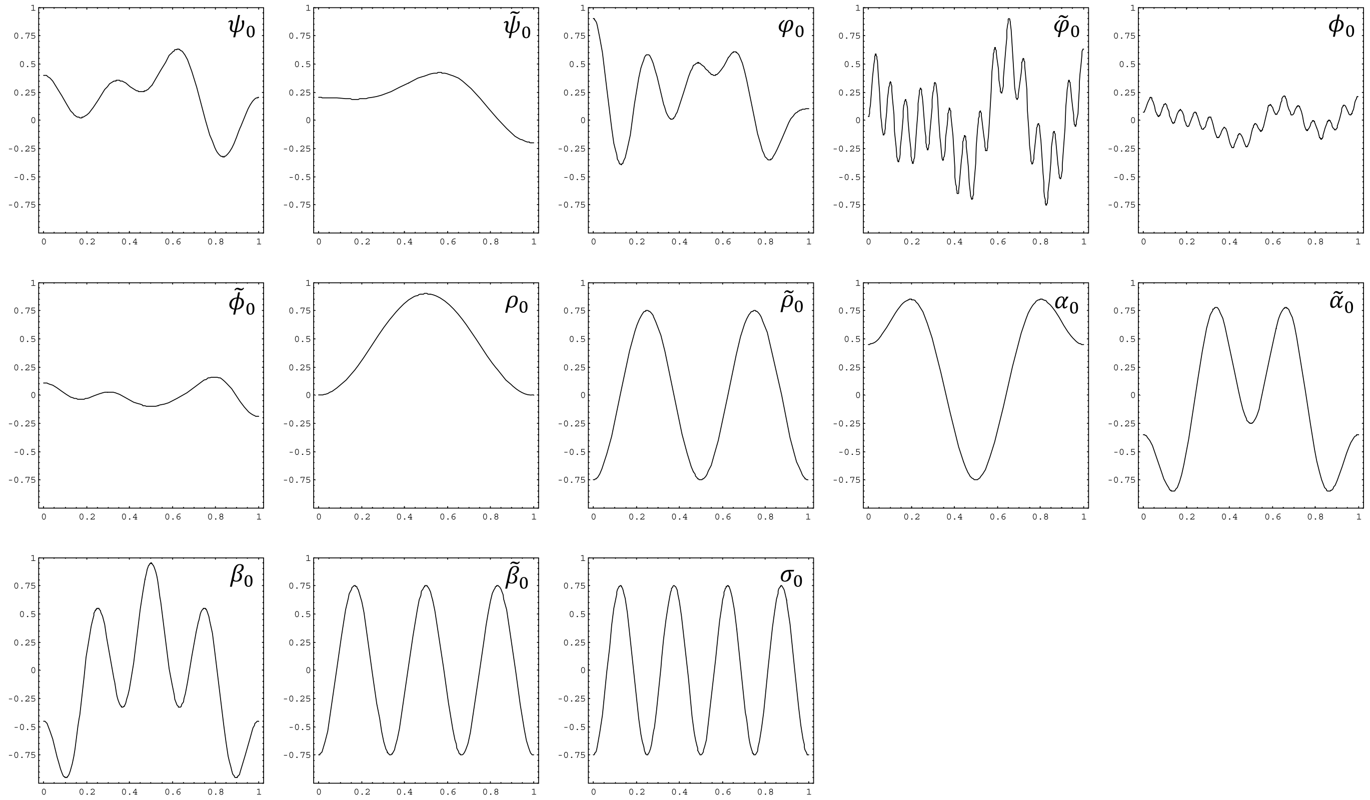


Fig. 7 Graphs of initial conditions ψ_0, $\widetilde{\psi}_0$, φ_0, $\widetilde{\varphi}_0$, ϕ_0, $\widetilde{\phi}_0$, ρ_0, $\widetilde{\rho}_0$, α_0, $\widetilde{\alpha}_0$, β_0, $\widetilde{\beta}_0$ and σ_0

4 Conclusions

The studied system is governed by a diffusion equation with a periodic nonlinear perturbation induced by an iterated map of the interval—a quadratic family f_μ. We have two parameters, the diffusion coefficient λ and the parameter characterizing f_μ, μ. The system exhibits a transient phase in which the number of critical points grows and its growth depends essentially on the topological entropy of f_μ. The dissipative nature of the diffusion equation eventually becomes dominant and the number of critical points stabilize around an average value, and with a certain dispersion. This average value, in the positive topological entropy case decays exponentially with the parameter λ and does not depend on μ. Moreover, the coefficients of the decay does not depend on the initial condition chosen. In a certain sense, the positive topological entropy produces an homogeneous exponential growth in the transient phase which disguise the differences in the initial conditions, normalizing the further behavior in the stable regime.

In the zero topological entropy, this does not happen and the significative differences in the initial conditions perpetuate through the stable regime. In this stable regime, the decay of the average number of critical points with λ and also its standard deviation depends strongly on the initial condition, as we can observe in the Table 3, for example, compare ψ_0 with $\widetilde{\phi}_0$, where $C_0\left(\psi_0\right) = 0.6368 \pm 0.0295$, $C_1\left(\psi_0\right) = -0.3266 \pm 0.0031$, and $C_0\left(\widetilde{\phi}_0\right) = 0.0053 \pm 0.0035$, $C_1\left(\widetilde{\phi}_0\right) = -0.7780 \pm 0.0615$.

Moreover, depending on μ, with zero topological entropy, we obtain periodical behavior for every initial condition, with the period given by the period of the critical point of f_μ. However, the periodic orbit itself is unique for each initial condition, which was somehow unexpected.

For the positive and zero topological entropy cases, we considered, for the coefficient λ, the intervals $[0.0000075, 0.01]$ and $[0.0000075, 0.0001]$, respectively, because the relevant dynamics occurs at these intervals (for example, for larger values of the coefficient λ, the system converges to a constant).

Acknowledgements This work has been partially supported by national funds by FCT—Funda-ção para a Ciência e a Tecnologia within the project UID/MAT/04674/2019.

References

1. Correia, M.F., Ramos, C.C., Vinagre, S.: The evolution and distribution of the periodic critical values of iterated differentiable functions. Nonlinear Anal. **75**, 6343–635 (2012)
2. Correia, M.F., Ramos, C.C., Vinagre, S.: Nonlinearly perturbed heat equation. Int. J. Pure Appl. Math. **94**(2), 279–296 (2014)
3. Milnor, J., Thurston, W.: On iterated maps of the interval. In: Alexander, J.C. (ed.) Proceedings University of Maryland, 1986–1987. Lecture Notes in Mathematics, vol. 1342, pp. 465–563. Springer, New York (1988)

4. Ramos, C.C., Santos, A.I., Vinagre, S.: A symbolic approach to nonlinearly perturbed heat equation. Int. J. Pure Appl. Math. **107**(4), 821–843 (2016)
5. Ramos, C.C., Santos, A.I., Vinagre, S.: Asymptotic behaviour in a certain nonlinearly perturbed heat equation: non periodic perturbation case. In: Proceedings of International Conference on Differential & Difference Equations and Applications 2017, pp. 581–593. Springer, Heidelberg (2018)
6. Romanenko, E.Yu., Sharkovsky, A.N.: From boundary value problems to difference equations: a method of investigation of chaotic vibrations. Int. J. Bifur. Chaos. **9**(7), 1285–1306 (1999)
7. Severino, R., Sharkovsky, A.N., Sousa Ramos, J., Vinagre, S.: Topological invariants in a model of a time-delayed Chua's circuit. Nonlinear Dyn. **44**, 81–90 (2006)
8. Sharkovsky, A.N.: Ideal turbulence. Nonlinear Dyn. **44**, 15–27 (2006)
9. Sharkovsky, A.N., Maistrenko, Yu., Romanenko, E.Yu.: Difference Equations and Their Applications. Kluwer Academic Publishers (1993)
10. Vinagre, S., Severino, R., Sousa Ramos, J.: Topological invariants in nonlinear boundary value problems. Chaos Solitons Fractals **25**, 65–78 (2005)

The Dynamics of a Hybrid Chaotic System

Carlos Ramos, Ana Isabel Santos, and Sandra Vinagre

Abstract We consider a forced damped piecewise linear oscillator whose motion is modeled by the second-order non-autonomous differential equation. Our hybrid chaotic system has a continuous regime, where the time flow is characterized by the explicit solutions of the ordinary differential equations, and a singular regime, where the time flow is characterized by an appropriate transformation linking the explicit solutions from one domain to the other. The purpose of this work is to study the complex behaviour of the system, namely the dependence on initial conditions and parameter variation.

1 Introduction

Piecewise linear dynamics may be used to study several mechanical systems such as gear box and rotor-bearing systems. For many years, the dynamics of gears has been of great interest to improve transmission and to reduce machinery noise. Although, in the initial phase, the linear vibration model developed provides a good prediction of gear vibration at low speeds, owing to high speed requirement in this type of systems, the linear vibration model is no longer adequate. So, in recent decades, with the aim of finding the origin of the vibration and noise, the piecewise linear model and the impact model were developed. In the literature, we find several models considering the piecewise linear system to describe engineering vibrations, such as vibration in gear box, rotor-bearing and elasto-plastic structures (see [1]). For example, in 1983, Shaw and Holmes [4] investigated a piecewise linear system with a single discontinuity using the mapping technique. More recently, Luo and Chen [1] presented an idealized

C. Ramos · A. I. Santos · S. Vinagre (✉)
Department of Mathematics, ECT, CIMA, University of Évora, Rua Romão Ramalho 59,
7000-671 Évora, Portugal
e-mail: smv@uevora.pt

C. Ramos
e-mail: ccr@uevora.pt

A. I. Santos
e-mail: aims@uevora.pt

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_48

piecewise linear system with impacts to model the vibration of gear transmission systems, which was investigated analytically through the corresponding mapping structures. Moreover, piecewise linear systems, on one hand have explicit solutions, since involves linear differential equations, on the other hand can be used to study chaotic nonlinear systems, through the methods we explain below.

In this paper, we consider a forced damped piecewise linear oscillator whose motion is modeled by the second-order non-autonomous differential equation

$$x'' + \alpha\, x' + g(x) = F \cos(\omega t), \tag{1}$$

where α is the damping coefficient, F is the forcing amplitude, ω is the forcing frequency and g is a linear function. Therefore, we have a continuous regime, where the time flow is characterized by the explicit solutions of the ordinary differential equations, and a singular regime, where the time flow is characterized by an appropriate transformation linking the explicit solutions from one domain to the other. In the continuous regime, we have in fact a linear regime. The phase space is partitioned in these continuous regimes, and in each set of the partition the system has a unique explicit solution, since the ODE is linear in each part. When the system is in a singular regime it changes to another region of the partition, entering again in the continuous regime. This method allow us to study a nonlinear system with very complex behaviour such as (1). Our differential dynamical system is studied by making use of numerical simulations, with similar techniques as the ones applied in [2].

From our previous work, [3], we know that the behaviour of the system, depending on the parameters, is simple, periodic or chaotic. The behaviour in the chaotic regime is characterized by phase-space trajectories exhibiting many orbits that are nearly closed. Moreover, in certain regions of the parameters there are sensitivity to the initial conditions and sensitivity to parameter perturbation. In the present work, we intend to investigate, making use of numerical simulations, where are these regions, namely we want determine the region where occurs periodic and chaotic motion and the existence of regions where the model explodes.

2 The Forced Damped Piecewise Oscillator Model

In this section, we present the problem whose behaviour we intended to study. In order to do that, consider that x represents the displacement, x' the velocity and x'' the acceleration, so the motion of a forced damped oscillator can be described by the second-order non-autonomous differential equation (1), where g is a linear piecewise function defined by

$$g(x) = (-1)^{j(x)} \frac{2}{\pi}\, x + (-1)^{j(x)+1} 2 j\,(x)\,,$$

where the function $j\left(x\right)$ is given by

$$j\left(x\right)=\frac{1}{\pi}\left(x+\frac{\pi}{2}\right)-\bmod\left[\frac{1}{\pi}\left(x+\frac{\pi}{2}\right),1\right].$$

The local solutions of Eq. (1) are known explicitly on each interval I_j, for $j\in\mathbb{Z}$, since the two families of differential equations involved are

$$x''+\alpha x'+\frac{2}{\pi}\,x-2j=F\cos(\omega t) \tag{2}$$

for $x\in I_j=\left[-\frac{\pi}{2}+j\pi,\frac{\pi}{2}+j\pi\right]$ and j even, and

$$x''+\alpha x'-\frac{2}{\pi}\,x+2j=F\cos(\omega t) \tag{3}$$

for $x\in I_j=\left[-\frac{\pi}{2}+j\pi,\frac{\pi}{2}+j\pi\right]$ and j odd.

So, as we can seen in [3], considering the initial conditions $x(t_0)=x_0\in I_j$, with j even, $x'(t_0)=v_0$ and $|\alpha|<\sqrt{\frac{8}{\pi}}$, the local solution of the family of differential equations (2) in each interval I_j, with j even, is

$$\begin{aligned} x(t)=e^{-\frac{\alpha}{2}(t-t_0)}&\left[A_1\cos\left(\sqrt{\beta_1}\,(t-t_0)\right)+A_2\sin\left(\sqrt{\beta_1}\,(t-t_0)\right)\right]\\ &+\frac{F\left(\frac{2}{\pi}-\omega^2\right)}{\alpha^2\omega^2+\left(\frac{2}{\pi}-\omega^2\right)^2}\cos\left(\omega\left(t-t_0\right)\right)\\ &+\frac{F\,\alpha\,\omega}{\alpha^2\omega^2+\left(\frac{2}{\pi}-\omega^2\right)^2}\sin\left(\omega\left(t-t_0\right)\right)+j\pi, \end{aligned} \tag{4}$$

where $\beta_1=\frac{2}{\pi}-\left(\frac{\alpha}{2}\right)^2$ and the coefficients A_1 and A_2, that depend on the initial conditions, are

$$A_1=x_0-j\pi-\frac{F\left(\frac{2}{\pi}-\omega^2\right)}{\left(\frac{2}{\pi}-\omega^2\right)^2+\alpha^2\omega^2},$$

$$A_2=-\frac{1}{\sqrt{\beta_1}}\left[\frac{F\,\alpha\,\omega^2}{\alpha^2\omega^2+\left(\frac{2}{\pi}-\omega^2\right)^2}-v_0+\frac{\alpha}{2}\left(j\pi-x_0+\frac{F\left(\frac{2}{\pi}-\omega^2\right)}{\alpha^2\omega^2+\left(\frac{2}{\pi}-\omega^2\right)^2}\right)\right].$$

On the other hand, the local solution of the family of differential equations (3) in each interval I_j, with j odd, based now on the initial conditions $x(t_0)=x_0\in I_j$, with j odd, and $x'(t_0)=v_0$, is given by

$$
\begin{aligned}
x(t) = e^{-\frac{\alpha}{2}(t-t_0)} \left[B_1\, e^{-\sqrt{\beta_2}(t-t_0)} + B_2\, e^{\sqrt{\beta_2}(t-t_0)} \right] + \\
- \frac{F\left(\frac{2}{\pi}+\omega^2\right)}{\alpha^2\omega^2 + \left(\frac{2}{\pi}+\omega^2\right)^2} \cos\left(\omega\left(t-t_0\right)\right) \\
+ \frac{F\,\alpha\,\omega}{\alpha^2\omega^2 + \left(\frac{2}{\pi}+\omega^2\right)^2} \sin\left(\omega\left(t-t_0\right)\right) + j\pi,
\end{aligned}
\tag{5}
$$

where $\beta_2 = \dfrac{2}{\pi} + \left(\dfrac{\alpha}{2}\right)^2$ and the coefficients B_1 and B_2 are

$$
\begin{aligned}
B_1 &= \frac{1}{2\sqrt{\beta_2}} \left[\frac{F\,\alpha\,\omega^2}{\alpha^2\omega^2 + \left(\frac{2}{\pi}+\omega^2\right)^2} - v_0 - \left(\sqrt{\beta_2} - \frac{\alpha}{2}\right) \right. \\
&\quad \left. \times \left(j\pi - x_0 - \frac{F\left(\frac{2}{\pi}+\omega^2\right)}{\alpha^2\omega^2 + \left(\frac{2}{\pi}+\omega^2\right)^2} \right) \right], \\
B_2 &= \left(j\pi - x_0 + \frac{F\left(\frac{2}{\pi}+\omega^2\right)}{\alpha^2\omega^2 + \left(\frac{2}{\pi}+\omega^2\right)^2} \right) \left(\frac{1}{2} + \frac{\alpha}{4\sqrt{\beta_2}} \right) + \\
&\quad + \frac{1}{2\sqrt{\beta_2}} \left(\frac{F\,\alpha\,\omega^2}{\alpha^2\omega^2 + \left(\frac{2}{\pi}+\omega^2\right)^2} - v_0 \right).
\end{aligned}
$$

Therefore, the families of solutions (4) and (5) can be systematically matched at

$$
x = -\frac{\pi}{2} + j\pi \text{ and } x = \frac{\pi}{2} + j\pi, \quad j \in \mathbb{Z},
$$

to obtain the global solution of the Eq. (1) as a continuous function.

In what follows, we present two examples to illustrate the behaviour of x as a function of time for several sets of parameters and initial conditions.

Example 1. Consider the damping coefficient $\alpha = 0.62$, the forcing frequency $\omega = 0.6$, the value of the forcing amplitude F between 1.08331 and 1.16367, and the initial conditions $x(0) = -1.00251$ and $x'(0) = 0$.

We can see in the Fig. 1 that the behaviour of the motion of the forced damped piecewise linear oscillator changes radically when the forcing amplitude F increases. Since, for the same values of the damping coefficient α and the forcing frequency ω, we obtain different types of periodic and aperiodic orbits, which exhibits different attractors.

Although the pattern presented in the Fig. 1(e) is not simple, it is not completely random. The behaviour in the chaotic regime is characterized by the phase-space trajectories exhibiting many orbits that are nearly closed. This is a common property of chaotic systems – they generally exhibit phase-space trajectories with significant structure.

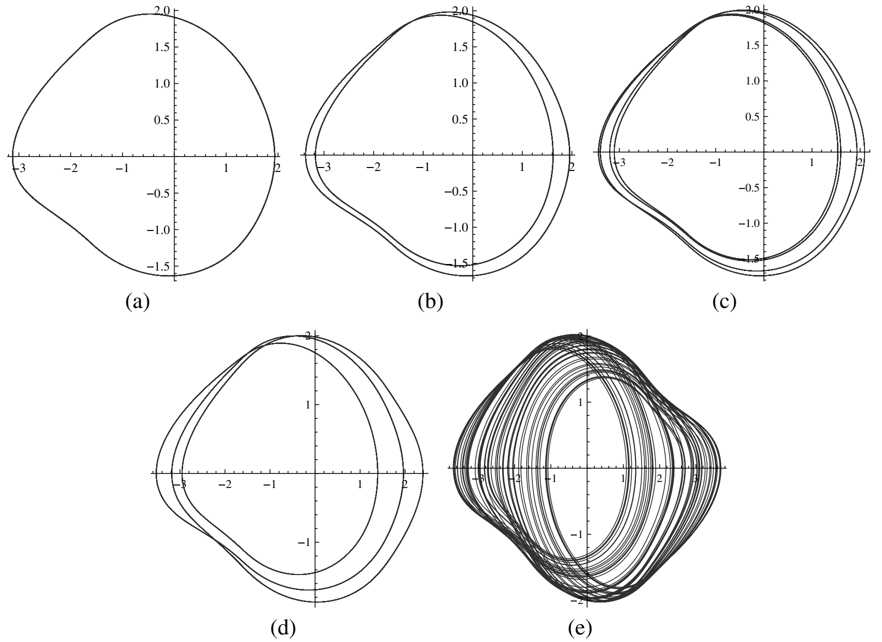


Fig. 1 Graphs of the orbits for **a** $F = 1.08331$, **b** $F = 1.12138$, **c** $F = 1.12984$, **d** $F = 1.14253$ and **e** $F = 1.16367$, with $\alpha = 0.62$, $\omega = 0.6$ and the initial conditions $x(0) = -1.00251$ and $x'(0) = 0$

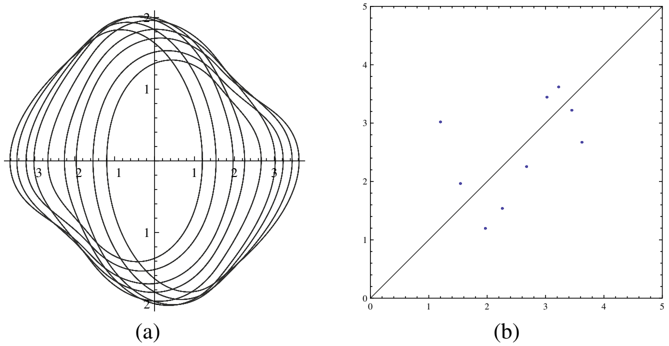


Fig. 2 Graphs of **a** the orbit and **b** the first return map of a periodic point, for $\alpha = 0.588$, $\omega = 0.6$, $F = 1.10681$ and the initial conditions $x(0) = 0.22$ and $x'(0) = 0$

Example 2. Consider the set of parameters $\alpha = 0.588$, $\omega = 0.6$ and $F = 1.10681$, and the initial conditions $x(0) = 0.22$ and $x'(0) = 0$.

In the Fig. 2, we show a periodic orbit of period nine and the correspondent first return map, which is plotted for the velocity of the forced damped piecewise linear oscillator and yields a chaotic behavior. The first return map indicates that, for this set of parameter values, the behaviour of x can be modeled by a one-dimensional iterated map.

3 Numeric Simulations

In this section, we present the numeric study of the behaviour of the model according to the variation of the parameter values and the initial conditions. Considering the value of the forcing frequency parameter ω fixed at 0.6, we studied the behaviour of x as a function of time for several sets of parameters of the damping coefficient α and the forcing amplitude F. The cases here presented correspond to four different types of initial conditions, namely we assume the motion of the forced damped piecewise linear oscillator evolves from a rest position, a non null position but with zero velocity, a non null speed and, finally, from a non null position and speed. In all these cases, we analyzed the existence of periodic and chaotic motion and the existence of regions where the model explodes, when the damping coefficient α varies between 0.45 and 1, in 0.02 step increments.

Note that in the next figures, which reflected the behaviour of the system for the cases previously refereed, to each colour corresponds a different period of the orbit (from lighter to darker), from the fixed point until period 2, 3, 4, ... and, finally, explosion.

The Rest Position Case

In the first numeric simulation, we analyze the behaviour of the motion of the forced damped piecewise oscillator when its starts from a rest position. Therefore, we consider the second-order non-autonomous differential equation (1) with the initial conditions $x(0) = 0$ and $x'(0) = 0$.

In the Fig. 3, we present the evolution of the system for the same values of the damping coefficient α when the forcing amplitude F increases. As we can see in this figure, in the region below the line $F_1(\alpha) \approx 0.1239 + 1.54744\alpha$ the system has only one critical point, which is a fixed point, regardless the parameter values for the damping coefficient and the forcing amplitude considered. However, the behaviour of the motion of the forced damped piecewise oscillator changes radically when the values of the forcing amplitude F increase above this line, since in the region between the lines $F_1(\alpha)$ and $F_2(\alpha) \approx 0.15 + 1.71\alpha$ we obtain different types of orbits ranging from periodic to aperiodic ones. On the order hand, in the region above the second line, $F_2(\alpha)$, the model always “explodes”.

The Non Null Position Case

Now we consider the cases where the motion of the forced damped piecewise oscillator starts with zero velocity, but the initial position is different of zero. In order to be able to establish a comparison with the previous example, we assumed the same values for the damping coefficient α and for the forcing amplitude F.

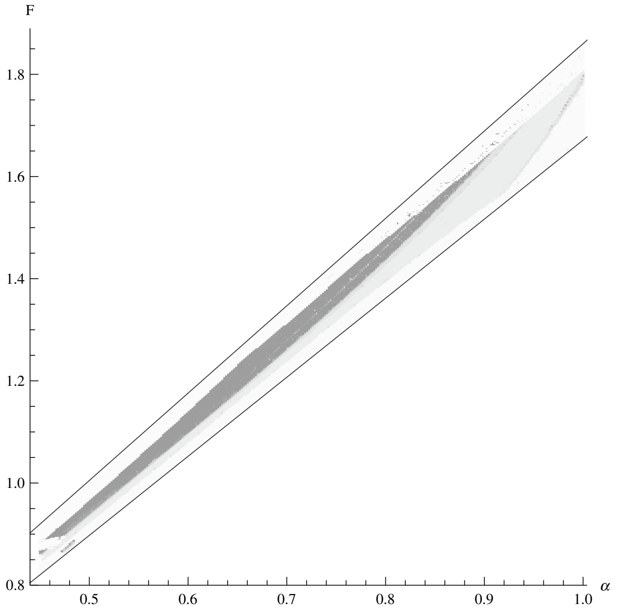


Fig. 3 Graph of the evolution of the period of the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = x'(0) = 0$

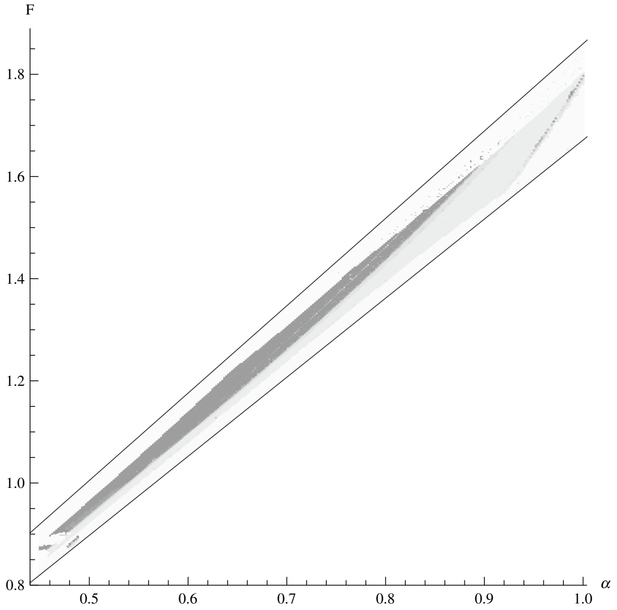


Fig. 4 Graph of the evolution of the period the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = 0.22$ and $x'(0) = 0$

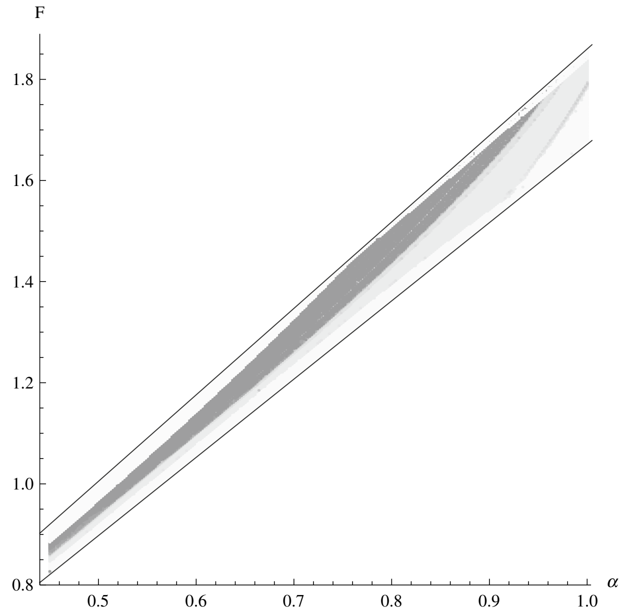


Fig. 5 Graph of the evolution of the period the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = -1.00251$ and $x'(0) = 0$

In the Figs. 4 and 5, we have the behaviour of the motion of the forced damped piecewise linear oscillator when we consider two different sets of initial conditions, $x(0) = 0.22$, $x'(0) = 0$ and $x(0) = -1.0025$, $x'(0) = 0$, respectively. In both cases, the pattern described above remains, which means that we have explosion in the region above the second line $F_2\ (\alpha)$ and a fixed critical point in the region below the first line $F_1\ (\alpha)$. In the region between the two lines, we have again major changes in the behaviour of the motion of the forced damped piecewise oscillator as the forcing amplitude F increases.

The Non Null Speed Case

The graphs of the simulations which follows were obtained considering that the system which describes the motion of the forced damped piecewise linear oscillator evolves from a non rest position, in both cases with non null initial velocity speed, that is, we consider the initial conditions $x(0) = 0$, $x'(0) = 0.25$ and $x(0) = 0$, $x'(0) = 0.715$.

For these cases, we consider again the values of the damping coefficient α between 0.45 and 1 and an increasing forcing amplitude F. So, in the graphs presented in the Figs. 6 and 7 we can see that the pattern described above remains, that is, we have a fixed critical point in the region below the line $0.1239 + 1.54744\alpha$ and explosion in the region above the line $0.15 + 1.71\alpha$. Once again, the evolution of the model changes quickly when the forcing amplitude F increases.

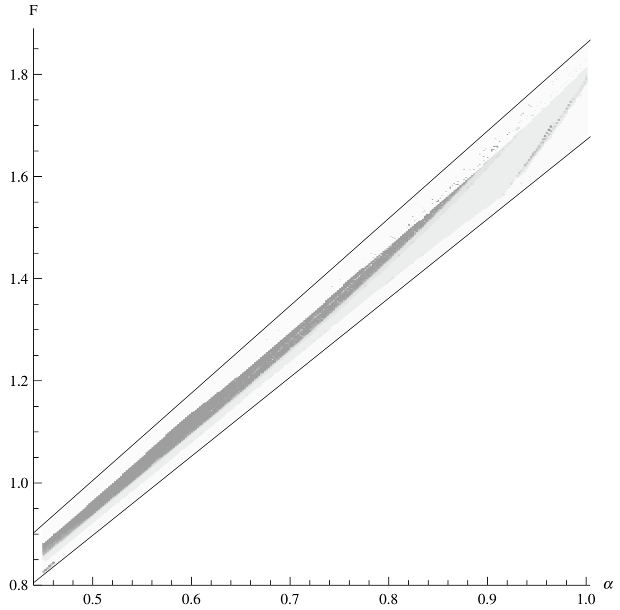


Fig. 6 Graph of the evolution of the period the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = 0$ and $x'(0) = 0.25$

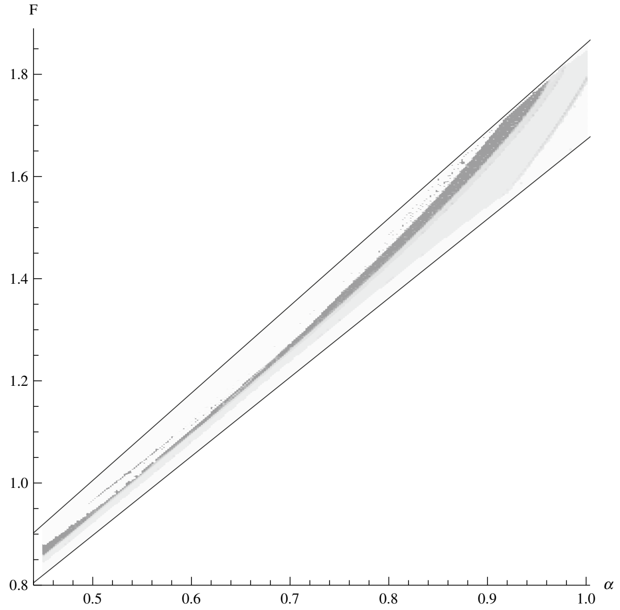


Fig. 7 Graph of the evolution of the period the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = 0$ and $x'(0) = 0.715$

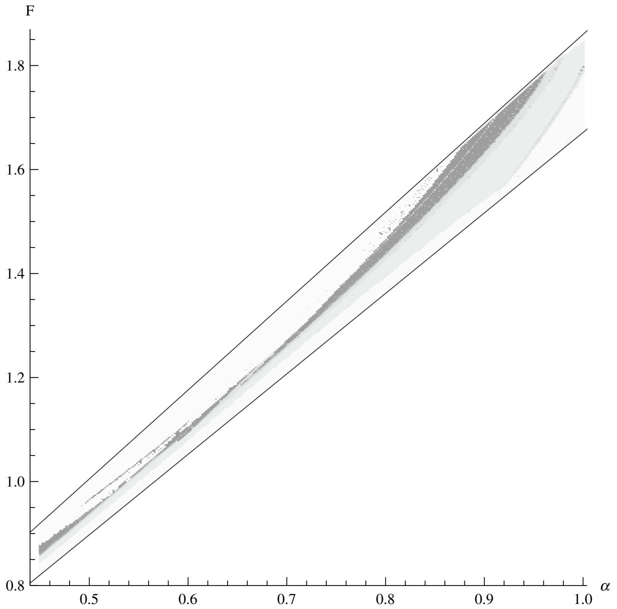


Fig. 8 Graph of the evolution of the period the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = 0.22$ and $x'(0) = 0.715$

The Non Null Position and Velocity Case

Finally, we present the case where the behaviour of the motion of the forced damped piecewise linear oscillator is analyzed assuming that it starts from a non rest position with non null position and speed.

In the Figs. 8 and 9, we present the graphs of the evolution of the motion where the initial conditions are, respectively, $x(0) = 0.22$, $x'(0) = 0.715$ and $x(0) = -1.00251$, $x'(0) = 0.25$, that is the case where the system evolves from a starting non null position with non null velocity. Comparing the simulations in both figures, we can see that, for higher values of α (greater than 0.88), the system with the first set of initial conditions reaches the explosion point later than the system with the second set of initial conditions. However, in both cases we have explosion in the region above the line $F_2(\alpha)$ and a fixed critical point in the region below the line $F_1(\alpha)$.

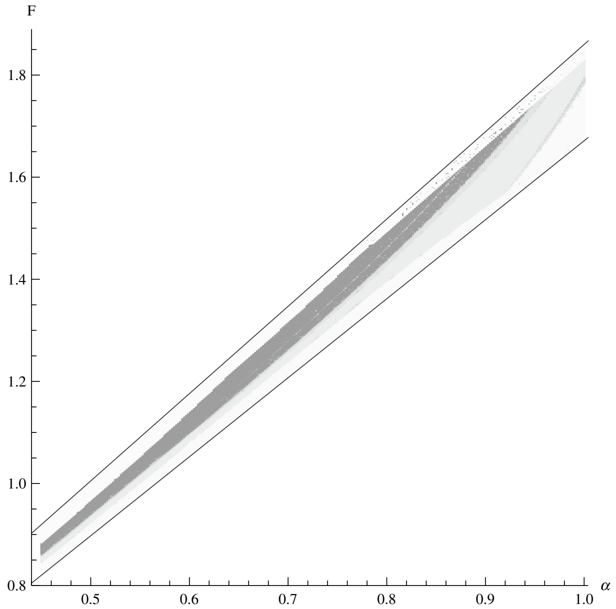


Fig. 9 Graph of the evolution of the period the orbits, when F increases and α varies between 0.45 and 1, with the initial conditions $x(0) = -1.00251$ and $x'(0) = 0.25$

4 Conclusions

The system analyzed is governed by a second-order non-autonomous differential equation that models a forced damped piecewise linear oscillator. Considering that the parameter corresponding to the forcing frequency ω is fixed at 0.6, our system has only two parameters, the damping coefficient α and the forcing amplitude F. So, we studied the behaviour of x as a function of time for several sets of parameters α and F and with different initial conditions, namely we analyzed the existence of periodic and chaotic motion and the existence of regions where the model explodes, for fixed values of the parameters. In all the cases analyzed, considering an increasing forcing amplitude F and that the values of the damping coefficient α varies between 0.45 and 1, we conclude that the system always has only one fixed critical point in the region below the line $F_1(\alpha) \approx 0.1239 + 1.54744\alpha$ and always explodes in the region above the second line $F_2(\alpha) \approx 0.15 + 1.71\alpha$. Nevertheless, when the parameters of the damping coefficient α and the forcing amplitude F take values in the region between theses two lines, that is, when $F_1(\alpha) \prec F(\alpha) \prec F_2(\alpha)$ the behaviour of the motion of the forced damped piecewise oscillator changes radically, since in this region we obtain different types of orbits ranging from periodic to aperiodic ones.

Acknowledgements This work has been partially supported by national funds by FCT—Fundação para a Ciência e a Tecnologia within the project UID/MAT/04674/2019.

References

1. Luo, A.C.J., Chen, L.: Periodic motions and grazing in a harmonically forced, piecewise, linear oscillator with impacts. Chaos Solit. Fractals **24**, 567–578 (2005)
2. Ramos, M.M., Ramos, C.C., Severino, R., Sousa Ramos, J.: Topological invariants of a chaotic pendulum. Int. J. Pure Appl. Math. **10**(2), 205–220 (2004)
3. Ramos, C.C., Santos, A.I., Vinagre, S.: Symbolic dynamics generated by a hybrid chaotic systems. Br. J. Math. Comput. Sci. **18**(2), 1–12 (2016)
4. Shaw, S.W., Holmes, P.J.: A periodically forced piecewise linear oscillator. J. Sound Vibr. **90**(1), 129–155 (1983)

On Regularity of Tetrahedral Meshes Produced by Some Red-Type Refinements

Sergey Korotov and Jon Eivind Vatne

Abstract In this work we propose a strategy for red-type refinements of tetrahedra which produces families of face-to-face tetrahedral partitions satisfying the maximum angle condition, a highly desired property in mesh generation, interpolation theory and finite element analysis.

1 Introduction

Along with various bisection-type methods (see e.g. the review [11]), the so-called red refinements (and associated with them red-green and green post-refinements) are the most popular techniques for refining simplicial meshes [1, 6, 10, 14–16], see Figs. 1 and 2.

However, the question of regularity of the meshes produced by the red-type technique is not always simple to answer as it is, in fact, not uniquely defined in dimensions three and higher, and moreover the similarity effect (always existing in the two-dimensional case) is present very rarely in higher dimensions, see Theorem 11. The situation leading to the non-uniqueness is illustrated in Fig. 2, where we see the red refinement technique applied to a tetrahedron (3D red refinement). It is easy to observe that there always exist three choices for selecting the interior diagonal inside of each father-tetrahedron to be refined as the diagonal should connect midpoints of two opposite edges of the tetrahedron considered. The situation is even harder in higher dimensions.

In this work we analyse one of the strategies for selecting such interior edges for tetrahedra involved and prove in a simple way that the resulting families of tetrahedral partitions satisfy the so-called maximum angle condition of Křížek (see (1)), using for that an equivalent form of this condition recently proposed in [8].

S. Korotov (✉) · J. E. Vatne
Department of Computer Science, Electrical Engineering and Mathematical Sciences,
Western Norway University of Applied Sciences, P.O. Box 7030, 5020 Bergen, Norway
e-mail: sergey.korotov@hvl.no

J. E. Vatne
e-mail: jon.eivind.vatne@hvl.no

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_49

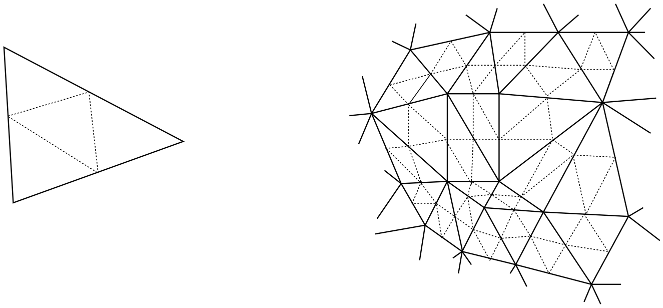
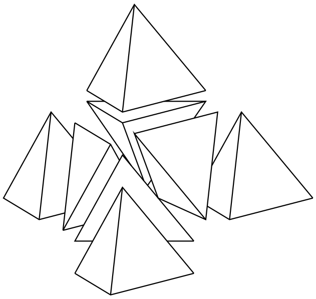


Fig. 1 Red refinements (using midpoints of edges) of a triangle and a triangulation

Fig. 2 3D red refinement (using midpoints of edges and midlines of triangular faces) of a tetrahedron



This regularity condition is still sufficient for guaranteeing the convergence of finite element aproximations [13] in various norms, and, at the same time, it presents a rather weak geometrical limitation so that it can be employed for real-life meshes with stretched elements, opposite to the so-called inscribed ball condition (see e.g. [3]), which prevents mesh elements from shrinking.

2 Mesh Regularity Definitions

Definition 1. A *polyhedron* is the closure $\overline{\Omega}$ of a bounded nonempty domain $\Omega \subset \mathbf{R}^3$ whose boundary can be expressed as a finite union of polygons.

Definition 2. A finite set of tetrahedra (denoted by the symbol T, possibly with subindices) is called a *(face-to-face or conforming) tetrahedral partition* of a polyhedron $\overline{\Omega}$ if

(i) the union of all the tetrahedra is $\overline{\Omega}$,
(ii) the interiors of the tetrahedra are mutually disjoint,
(iii) any face of any tetrahedron from the set is either a face of another tetrahedron in the set, or a subset of $\partial\Omega$.

Theorem 3. *For any polyhedron there exists a conforming partition into tetrahedra.*

The constructive proof is presented e.g. in [12].

Definition 4. For a given tetrahedral partition $\mathcal{T}_h$ the *discretization parameter* h (called also the *mesh size*) stands for the maximum length of all edges in the partition, i.e.,

$$h = \max_{T \in \mathcal{T}_h} h_T,$$

where

$$h_T = \operatorname{diam} T.$$

To prove various convergence statements in numerical analysis, we usually work with infinite sequences of partitions when the associated discretization parameter tends to zero.

Definition 5. A set of partitions $\mathcal{F}$ is called a *family of partitions* if for every $\varepsilon > 0$ there exists $\mathcal{T}_h \in \mathcal{F}$ such that $h < \varepsilon$.

The following maximum angle conditions for tetrahedral meshes was proposed by Křížek in [13]: there exists a constant $\gamma_0 < \pi$ such that for any face-to-face tetrahedralization $\mathcal{T}_h \in \mathcal{F}$ and any tetrahedron $T \in \mathcal{T}_h$ one has

$$\gamma_{\mathrm{D}} \le \gamma_0 \quad \text{and} \quad \gamma_{\mathrm{F}} \le \gamma_0, \tag{1}$$

where γ_{D} is the maximum dihedral angles between faces of T and γ_{F} is the maximum angle in all four triangular faces of T.

Remark 6. The maximum angle condition (1) covers several types (needle, splinter, and wedge) of tetrahedral degeneracies illustrated in Fig. 3. Its natural analog to higher dimensions is presented in [9].

In 1978, Eriksson introduced a generalization of the sine function to an arbitrary d-dimensional spatial angle, see [5, p. 74], which reads for the tetrahedral case as follows.

Definition 7. Let $\hat{A}_i$ be the spatial angle at the vertex A_i, $i = 0, 1, 2, 3$, of the tetrahedron $T = \operatorname{conv}\{A_0, A_1, A_2, A_3\}$. Then 3-sine of the angle $\hat{A}_i$ is computed as

$$\sin_3(\hat{A}_i | A_0 A_1 A_2 A_3) = \frac{9\,(\operatorname{meas}_3 T)^2}{2\,\Pi_{j=0, j \ne i}^{3} \operatorname{meas}_2 F_j}, \tag{2}$$

where F_j is the triangular face opposite to the vertex A_j.

Remark 8. The above definition can be applied to three (linearly independent) vectors by letting A_i be the point of origin of the vectors and the remaining vertices be the endpoints of the vectors. In this formulation, Eriksson proves that the value of $\sin_3$ is unchanged if the vectors are rescaled by a nonzero factor (positive or negative).

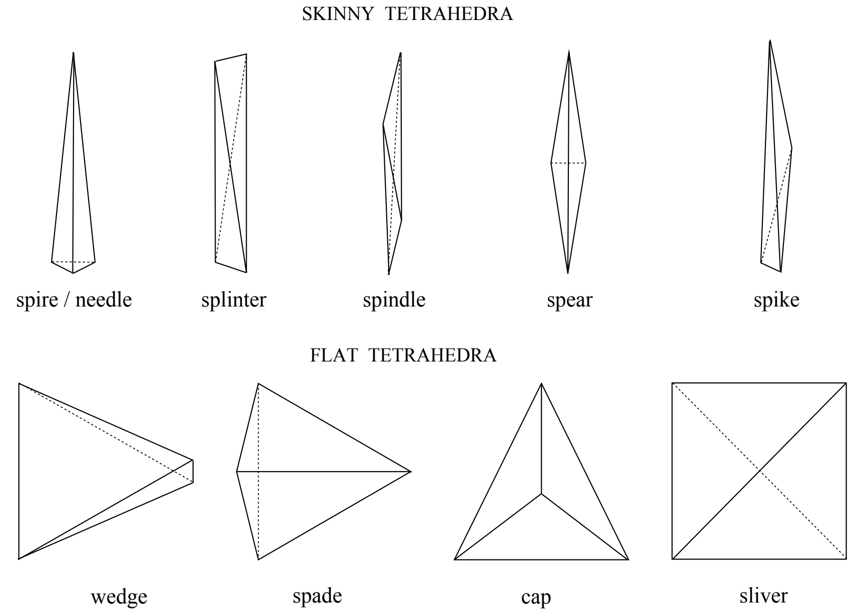


Fig. 3 Classification of degenerated (skinny and flat) tetrahedra according to [2, 4]

Definition 9. A family $\mathcal{F} = \{\mathcal{T}_h\}_{h\to 0}$ of face-to-face tetrahedral partitions of a polyhedron $\overline{\Omega}$ is said to satisfy the *generalized maximum angle condition* if there exists a constant $C > 0$ such that for any $\mathcal{T}_h \in \mathcal{F}$ and any $T \in \mathcal{T}_h$ one can always choose 3 edges of T, which, when considered as vectors, constitute a (higher-dimensional) angle whose 3-sine is bounded from below by the constant C.

In this definition, we actually allow both possible scenaria: the three edges emanate from a common vertex (forming a corner), or they go from vertex to vertex throughout the tetrahedron (forming a path).

Theorem 10. *The generalized maximum angle condition of Definition 9 is equivalent to the maximum angle condition (1).*

For the proof see [9].

3 Main Result

First, we mention several auxiliary results closely related to the regularity of tetrahedral meshes produced by the red-type refinements.

Theorem 11. *There exists only one type of tetrahedron T (up to similarity) whose red refinement produces eight congruent subtetrahedra similar to T. It is called the Sommerville tetrahedron T_1.*

The proof is given in [10, 16].

The next result immediately follows from the fact that the four "exterior" subtetrahedra arising from the red refinement algorithm are similar to the original tetrahedron, see Fig. 4 (right).

Theorem 12. *The maximum (minimum) dihedral angles between faces and also the maximum (minimum) angles in all triangular faces of all tetrahedra $T \in \mathcal{T}_h \in \mathcal{F}$ generated by the red-type refinements form nondecreasing (nonincreasing) sequences as $h \to 0$.*

Red Refinement Algorithm: In a given tetrahedron we fix three edges forming a path, see Fig. 4 (left), where these edges are marked by bold lines and have lengths $2a$, $2b$, and $2c$. For the red refinement to be performed we select the only interior edge which connects two opposite edges which both are not among those three selected ones. The resulting partition is presented in Fig. 4 (right).

Main Properties of the Algorithm: A simple case-by-case analysis shows that each of eight resulting subtetrahedra has a path consisting of halves of those three edges (i.e. a, b, and c) selected in the father tetrahedron (however, not necessarily in the same order). See again Fig. 4 for an illustration. Moreover, all edges marked there by a are parallel to the edge $2a$, and similarly all edges marked by b are parallel to the edge $2b$, and all edges marked by c are parallel to $2c$. We choose this path to repeat the red refinement.

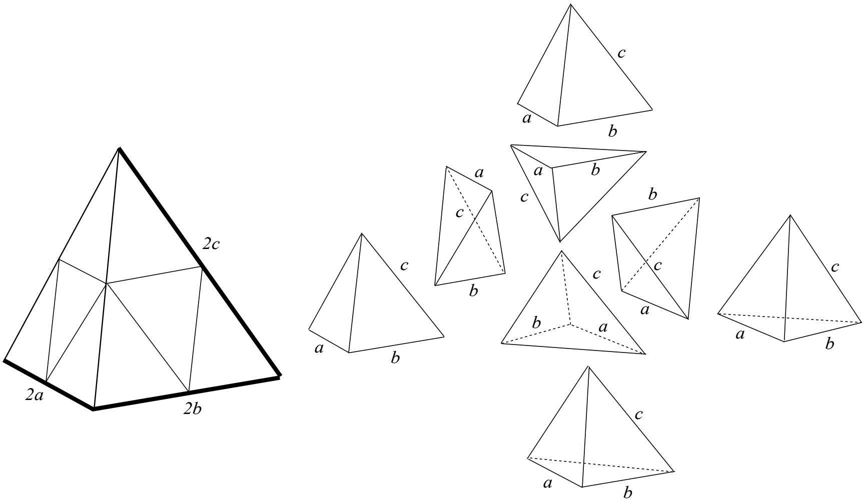


Fig. 4 3D red refinement explosed

The main result of the paper reads now as follows.

Theorem 13. *Given a face-to-face tetrahedral partition $\mathcal{T}_{h_0}$ of a polyhedron $\overline{\Omega}$. Then the family of partitions $\mathcal{F} = \{\mathcal{T}_h\}_{h\to 0}$, generated from $\mathcal{T}_{h_0}$ by the red refinement algorithm presented above, satisfies the generalized maximum angle condition of Definition 9.*

Proof. For any selections of interior edges in refined tetrahedra (at any refinement level), the faces of all tetrahedra are split in the same way, so there is no problem with guaranteeing overall conformity.

Now, in each tetrahedron T from $\mathcal{T}_{h_0}$ we fix a path of three edges $2a_T$, $2b_T$, $2c_T$ and calculate three-dimensional sine of the angle made by these edges. From the properties of the algorithm discussed in above we observe that at each refinement level for any resulting subtetrahedra we always have a triple of edges which are parallel to those initial fixed three edges in the father tetrahedron from $\mathcal{T}_{h_0}$. Using Remark 8, we also observe that $\sin_3$ of all these triples are just the same as $\sin_3$ of the initial fixed edges in the father tetrahedron. And now the validity of the generalized maximum angle condition immediately follows with a constant

$$C = \min_{T\in\mathcal{T}_{h_0}} \sin_3(2a_T, 2b_T, 2c_T).$$

4 Final Comments

1. It would be interesting to get similar results in higher dimensions.
2. The maximum angle condition is only a sufficient condition for FEM convergence, as demonstrated in [7], so it would be important to find some weaker regularity concepts especially taking into account that not all degeneracies of Fig. 3 are covered by this maximum angle condition.

References

1. Bey, J.: Simplicial grid refinement: on Freudenthal's algorithm and the optimal number of congruence classes. Numer. Math. **85**, 1–29 (2000)
2. Cheng, S.W., Dey, T.K., Edelsbrunner, H., Facello, M.A., Teng, S.H.: Sliver exudation. In: Proceedings of the 15th ACM Symposium on Computational Geometry, pp. 1–13 (1999)
3. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
4. Edelsbrunner, H.: Triangulations and meshes in computational geometry. Acta Numer. **9**, 133–213 (2000)
5. Eriksson, F.: The law of sines for tetrahedra and n-simplices. Geom. Dedicata. **7**, 71–80 (1978)
6. Grande, J.: Red-green refinement of simplicial meshes in d dimensions. Math. Comput. **88**, 751–782 (2019)

7. Hannukainen, A., Korotov, S., Křížek, M.: The maximum angle condition is not necessary for convergence of the finite element method. Numer. Math. **120**, 79–88 (2012)
8. Hannukainen, A., Korotov, S., Křížek, M.: Generalizations of the Synge-type condition in the finite element method. Appl. Math. **62**, 1–13 (2017)
9. Khademi, A., Korotov, S., Vatne, J.E.: On the generalization of the Synge-Křížek maximum angle condition for d-simplices. J. Comput. Appl. Math. **358**, 29–33 (2019)
10. Korotov, S., Křížek, M.: Red refinements of simplices into congruent subsimplices. Comput. Math. Appl. **67**, 2199–2204 (2014)
11. Korotov, S., Plaza, Á., Suárez, J.: Longest-edge n-section algorithms: properties and open problems. J. Comput. Appl. Math. **293**, 139–146 (2016)
12. Křížek, M.: An equilibrium finite element method in three-dimensional elasticity. Aplikace Matematiky **27**, 46–75 (1982)
13. Křížek, M.: On the maximum angle condition for linear tetrahedral elements. SIAM J. Numer. Anal. **29**, 513–520 (1992)
14. Křížek, M., Strouboulis, T.: How to generate local refinements of unstructured tetrahedral meshes satisfying a regularity ball condition. Numer. Methods Partial Differ. Equ. **13**, 201–214 (1997)
15. Ong, M.E.G.: Uniform refinement of a tetrahedron. SIAM J. Sci. Comput. **15**, 1134–1144 (1994)
16. Zhang, S.: Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. Houston J. Math. **21**, 541–556 (1995)

Difference Scheme for Partial Differential Equations of Fractional Order with a Nonlinear Differentiation Operator

Svyatoslav Solodushkin, Tatiana Gorbova, and Vladimir Pimenov

Abstract A fractional differential equation in partial derivatives with non-linearity in differentiation operator is considered. We developed a numerical method which has the second order of convergence in time and first order in space and could be considered as a fractional analog of Crank–Nicolson method. Nonlinear high dimensional systems which arise on each time layer are solved iteratively. The method is proven to be consistent and unconditionally stable. Results of numerical examples coincides with theoretical ones.

1 Introduction

Fractional differential equations have found interesting applications in many fields of the natural sciences and engineering, including the theory of viscoelasticity, the theory of thermoelasticity, financial problems, self-similar protein dynamics and population dynamics, see [2, 7, 8] and loads of references therein. Since many natural processes are nonlinear, we are dictated to consider fractional differential equations in partial derivatives with non-linearity in the differentiation operators. From the point of view of computational mathematics these equations is an exceptionally complex, fascinating and little-studied object.

Explicit solution of such equations could be found in exceptional cases only, therefore the elaboration, substantiation and program realization of numerical methods for these equations are of great interest.

Numerical methods for partial fractional differential equations where the non-linearity could be involved in a heterogeneous function, but not in the differential operators, have been elaborated and studied in the past decades. Below we review some approaches to their numerical solving.

S. Solodushkin (✉) · T. Gorbova · V. Pimenov
Ural Federal University, Yekaterinburg, Russia
e-mail: s.i.solodushkin@urfu.ru

T. Gorbova
e-mail: tvgorbova@gmail.com

V. Pimenov
e-mail: v.g.pimenov@urfu.ru

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_50

There are two basic types of fractional equations [7]: with fractional derivatives in time, where the Caputo definition is mainly used, and fractional derivatives in space, where the Riemann–Liouville and the Riesz definitions are used. In this paper we consider the second type only.

In [9] an approach based on the classical Crank–Nicholson method was used to solve initial-boundary value fractional diffusive equations. Stability, consistency, and convergence were examined. It was shown that the fractional analog of the Crank–Nicholson method based on the shifted Grünwald formula is unconditionally stable. The Richardson extrapolation method was used to increase the convergence order with respect to space up to the second.

Implicit difference schemes for fractional partial differential equations with time delay were constructed in [4, 5]. The authors used shifted Grünwald–Letnikov formulas for the approximation of fractional derivatives with respect to spatial variables and the L1-algorithm for the approximation of fractional derivatives in time. The technique is very similar to used in the present work.

Energy-preserving finite-difference scheme with fractional centered differences were presented in [2].

At the same time numerical methods for partial fractional differential equations with a non-linearity in the differentiation operators have not been studied yet. In [8], as in most similar works, numerical methods are not considered, but attempts are made to find the exact solution in the form of series.

The elaboration of difference schemes for partial fractional differential equations with a non-linearity in the differentiation operators is associated with a number of difficulties. Numerical experiments showed that the explicit schemes lead to instability. On the other hand direct application of the implicit scheme leads to the necessity to solve nonlinear systems of large dimension. This work is a continuation of [1] where initial-boundary value problem in partial derivatives with a non-linearity in the differentiation operators was considered for integer-order case.

In this paper we consider an equation of the following form

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial^\alpha \phi(p(x,t))}{\partial x^\alpha} + f(x,t), \tag{1}$$

where t and x are independent variables, $0 \le t \le T,\ 0 \le x \le X$, and $p(x,t)$ is an unknown function to be found, ϕ is a given non-linear function. The left-sides fractional derivative is defined in Riemann–Liouville sense

$$\frac{\partial^\alpha F(x)}{\partial x^\alpha} = \frac{1}{\Gamma(n-\alpha)} \frac{d^n}{dx^n} \int_0^x \frac{F(\xi)}{(x-\xi)^{\alpha-n+1}}\, d\xi,$$

where n is integer such that $n-1 < \alpha \le n$, and it is supposed that $F(x) = 0$ for $x \le 0$. In the rest of this article we consider the case $1 < \alpha \le 2$.

Initial and boundary conditions are set as follow

$$p(x,0) = \varphi(x),\ 0 \le x \le X, \tag{2}$$

$$p(0,t) = p_0(t),\ p(X,t) = p_1(t),\ 0 \le t \le T. \tag{3}$$

Note that $\alpha = 2$ is the classical diffusion equation. The case of $1 < \alpha \le 2$ models a super-diffusive flow in which a cloud of diffusing particles spreads at a faster rate than the classical diffusion model predicts, and $\alpha = 1$ corresponds to the classical advective flow.

In this paper the technique described in [10] is used: by means of the change of variables, the non-linearity in the differentiation operator with respect to the spatial variable is transmitted to the time differentiation operator. Then an implicit difference scheme is constructed, the appeared nonlinear system is solved by the Newton method. The main result consists in proving the stability and convergence of the constructed algorithm.

2 Implicit Difference Scheme

Assuming the single-valued invertibility of $\phi(p)$ on the domain of our interest, we make the substitution $u = \phi(p),\ p = \omega(u)$, then (1) is transformed to the form

$$\frac{\partial \omega(u(x,t))}{\partial t} = \frac{\partial^\alpha u(x,t)}{\partial x^\alpha} + f(x,t), \tag{4}$$

and the initial and boundary conditions could be rewritten as follow:

$$u(x,0) = \phi(\varphi(x)),\ 0 \leqslant x \leqslant X, \tag{5}$$

$$u(0,t) = \phi(p_0(t)) = \mu_0(t),\ u(X,t) = \phi(p_1(t)) = \mu_1(t),\ 0 \le t \le T. \tag{6}$$

We shall assume that the problem (4)–(6) has a unique solution, understood in the classical sense, and this solution has continuous derivatives with respect to state variables x up to fourth order, continuous derivatives with respect to time t up to second order. Also, we assume that ω is twice continuously differentiable in its domain and its first derivative is uniformly greater than zero

$$0 < \hat{\omega} \le \omega'(u). \tag{7}$$

We consider an equidistant partition of $[0, X]$ into parts with step size $h = X/N$ and define the grid $x_i = ih,\ i = 0, \dots, N$. We also split the time interval $[0, T]$ into M parts with step size $\varDelta = T/M$ and define the grid $t_j = j\varDelta$, $j = 0, \dots, M$.

Denote by u_j^i the approximation of the function value $u(x_i, t_j),\ i = 0, 1, \dots N$, $j = 0, \dots M$, at the respective node.

To approximate the left-sides fractional derivative in the internal grid nodes we use the shifted Grünwald formula [3]

$$\frac{\partial^\alpha u(x_i,t_j)}{\partial x^\alpha} \approx \frac{1}{h^\alpha}\sum_{s=0}^{i+1} g_{\alpha,s}\, u(x_{i+1-s},t_j), \quad 1 \le i \le N-1,$$

where the normalized Grünwald weights are defined as follow $g_{\alpha,0} = 1$ and $g_{\alpha,s} = (-1)^s \, \frac{\alpha(\alpha-1)\ldots(\alpha-s+1)}{s!}, \ s = 1, 2, 3, \ldots$.

Consider a nonlinear implicit difference scheme, $j = 0, 1, \ldots, M-1$,

$$\begin{aligned} \frac{\omega(u^i_{j+1}) - \omega(u^i_j)}{\Delta} = \frac{1}{2h^\alpha}\left(\sum_{s=0}^{i+1} g_{\alpha,s}\, u^{i+1-s}_{j+1} + \sum_{s=0}^{i+1} g_{\alpha,s}\, u^{i+1-s}_{j}\right) + f^i_{j+1/2}, \\ \text{for } i = 1, \ldots, N-1, \\ \text{and } u^0_{j+1} = \mu_0(t_{j+1}),\ u^N_{j+1} = \mu_1(t_{j+1}) \end{aligned} \tag{8}$$

with initial conditions $u^i_j = \phi(\varphi(x_i)),\ i = 0, \ldots, N$. To make notation shorter $f^i_{j+1/2}$ denotes $f(x_i, t_j + \Delta/2)$.

For each fixed j the (8) is a system of equations that are nonlinear with respect to $u^i_{j+1},\ i = 1, \ldots, N-1$. To solve (8) at each time layer j we apply Newton's method [10],

$$\begin{aligned} &\omega(u^i_{j+1}[k]) + \omega'(u^i_{j+1}[k])(u^i_{j+1}[k+1] - u^i_{j+1}[k]) - \omega(u^i_j) \\ &= \frac{\Delta}{2h^\alpha}\left(\sum_{s=0}^{i+1} g_{\alpha,s}\, u^{i+1-s}_{j+1} + \sum_{s=0}^{i+1} g_{\alpha,s}\, u^{i+1-s}_{j}\right) + \Delta f^i_{j+1/2}, \end{aligned} \tag{9}$$

where k is an iteration number, $k = 0, 1, \ldots$, and $u^i_{j+1}[k]$ is $k-$th approximation by the Newton's method to $u^i_{j+1},\ i = 1, \ldots, N-1$. Note, when we search $u^i_{j+1}[k+1]$ in (9) we use u^i_j which represents not the exact solution obtained at the j-th time layer (it is actually unknown) but its approximation in Newton's method.

Let us denote $y_j = (u^1_j, u^2_j, \ldots, u^{N-1}_j)^T \in Y$, where Y is a vector space of dimension $N-1$ and T is a transpose sign.

Let us consider a matrix A which elements $A_{i,j}$ are defined as follow

$$A_{i,j} = \begin{cases} \eta\, g_{\alpha,i-j+1} & \text{for } j \le i-1 \\ \eta\, g_{\alpha,1} & \text{for } j = i \\ \eta\, g_{\alpha,0} & \text{for } j = i+1 \\ 0 & \text{for } j > i+1 \end{cases},$$

where $\eta = \dfrac{1}{2h^\alpha}$. We also define vector functions $\omega(y_j)$ and f_j as vectors with components $\omega(u^i_j)$ and $f(x_i, t_j)$ respectively, then system (8) could be represented as follow

$$\omega(y_{j+1}) - \Delta A y_{j+1} = \Delta A y_j + \omega(y_j) + \Delta f_{j+1/2}. \tag{10}$$

In a similar way we denote $y_j[k] = (u_j^1[k], u_j^2[k], \dots, u_j^{N-1}[k])^T \in Y$, also we denote by $\omega'(y_j)$ the diagonal matrix with $\omega'(u_j^i)$ on the main diagonal in i-th row.

Lemma 1. *[3] Matrix $\omega'(y_{j+1}[k]) - \Delta A$ is positively defined.*

According to Lemma 1 the linear system, which should be solved at each iteration of (9), is solvable for each k, and then the iterative process (9) with exactly K iterations on each time layer can be written in the form

$$\begin{aligned} y_{j+1}[k+1] = (\omega'(y_{j+1}[k]) - \Delta A)^{-1} \times \\ \times \Big(\omega(y_j[K]) + \Delta A y_j[K] + \omega'(y_{j+1}[k]) y_{j+1}[k] - \omega(y_{j+1}[k])\Big) + \\ + \Delta(\omega'(y_{j+1}[k]) - \Delta A)^{-1} f_{j+1/2}, \quad k = 0, \dots, K-1, \end{aligned} \tag{11}$$

$$y_{j+1}[0] = y_j[K]. \tag{12}$$

Each iteration involves the need to calculate the value of a nonlinear (with respect to $y_{j+1}[k]$) operators and solve a linear (with respect to $y_{j+1}[k+1]$) system. Since $\omega'(y_{j+1}[k]) - \Delta A$ is an almost triangular matrix, it could be effectively inverted using special algorithms. Therefore the iterative algorithm is not computationally expensive.

To study the convergence of difference scheme (8) supplemented by iterative method (11)–(12) we will look at them from the point of view of functional analysis and operator equations. To do this, we present in the next section the corresponding results on the convergence of nonlinear difference schemes in general form [6]. After that we embed a method (8), (11)–(12) into the general scheme.

3 Convergence of General Nonlinear Difference Schemes

Let a segment $[0, T]$ be divided into M parts with step $\Delta = T/M$ and nodes are defined as $t_j = j\Delta,\ j = 0, \dots, M$. A discrete model is defined as a grid function $y_j = y(t_j) \in Y,\ \ j = 0, \dots, M$, where Y is q-dimensional normed space with norm $\|\cdot\|_Y$. We will assume that the dimension q of the space Y depends on some parameter $h > 0$. Note, that in the previous section h corresponded to the grid step with respect to space.

Initial value of the model is defined as $y(t_0) = y_0$.

The formula of advance of the model by a step is, by definition, the relation

$$y_{j+1} = S(y_j) + \Delta \Phi(y_j), \tag{13}$$

where the transition operator $S(y_j) = S(y_j; t_j, \Delta, h)$ is a nonlinear operator which is Lipschitz with a constant $L_S = L_S(\Delta, h)$ and $\Phi(y_j) = \Phi(y_j; t_j, \Delta, h)$ is a nonlinear operator which is Lipschitz with a constant $L_\Phi = L_\Phi(\Delta, h)$.

The function of exact values is, by definition, the mapping

$$Z(t_j, \Delta, h) = z_j \in Y, \quad j = 0, \dots, M.$$

To know the function of exact values means to know the exact solution of the original problem in the nodes. In what follows, for simplicity, we assume that the initial value coincide with the initial value of the function of the exact values $y_0 = z_0$.

We will say that the method (13) converges if there exists a constant C independent of Δ and h and a function $q(\Delta, h)$, $\lim\limits_{\Delta \to 0, h \to 0} q(\Delta, h) = 0$, such that the following inequality holds:

$$\| z_j - y_j \|_Y \leq Cq(\Delta, h) \tag{14}$$

for all $j = 0, \dots, M$. Function $q(\Delta, h)$ defines the order of convergence.

The order of convergence depends on the approximation error and the stability properties of the method. An error of approximation (a residual) is, by definition, the grid function

$$d_j = (z_{j+1} - S(z_j))/\Delta - \Phi(z_j), \;\; j = 0, \dots, M - 1. \tag{15}$$

We will say that the error of approximation in method (13) has an order of $q(\Delta, h)$, if there exists a constant C independent of Δ and h such that for all $j = 0, \dots, M - 1$ the following inequality holds:

$$\| d_j \|_Y \leq Cq(\Delta, h).$$

The method (13) is said to be stable, if

$$L_S = L_S(\Delta, h) \leq 1. \tag{16}$$

Theorem 1. *Let method (13) is stable, the error of approximation has an order $q(\Delta, h)$, $\lim\limits_{\Delta \to 0, h \to 0} q(\Delta, h) = 0$, then method (13) converges with order $q(\Delta, h)$.*

Proof. Let us denote $\delta_j = \| z_j - y_j \|_Y$, $j = 0, \dots, M$. Since S and Φ are Lipschitz ones, for all $j = 0, \dots, M - 1$ we have

$$\delta_{j+1} = \| S(z_j) + \Delta\Phi(z_j) + \Delta d_j - S(y_j) - \Delta\Phi(y_j) \|_Y \leqslant L_S\delta_j + \Delta L_\Phi \delta_j + \Delta \| d_j \|_Y .$$

Using the stability condition, we obtain

$$\delta_{j+1} \leqslant (1 + \Delta L_\Phi)\delta_j + \Delta \| d_j \|_Y .$$

Taking into account that $\delta_0 = 0$, from this estimate the following estimate could be deduced by standard methods (see for example [6])

$$\delta_j \leqslant \frac{D}{L_\Phi}\, exp(TL_\Phi),$$

where $D = \max\limits_{0 \le l \le M} \|d_l\|_Y$. The last estimate is right for all $j = 0, \dots, M$. This implies the conclusion of the theorem. □

The transformation of the nonlinear difference scheme to explicit form (13) is usually not an easy task. So, let us consider the approximation of this scheme in the form of an iterative process[1]

$$y_{j+1}[k+1] = \tilde{S}(y_{j+1}[k]; y_j[K]) + \Delta\tilde{\Phi}(y_{j+1}[k]; y_j[K]),\ k = 0, \dots, K-1, \tag{17}$$

where the number of iterations K on each time layer j is fixed, and as the initial approximation on each time layer it is taken $y_{j+1}[0] = y_j[K]$. Iterative process (17) is reduced to the form

$$y_{j+1}[K] = \hat{S}_K(y_j[K]) + \Delta\hat{\Phi}_K(y_j[K]). \tag{18}$$

The definitions introduced above for the general nonlinear scheme (13) could be transformed in an obvious way for the *approximation scheme* (18).

We will say that the method (18) converges if there exists a constant C and a function $q(\Delta, h, K)$,

$$\lim_{\Delta \to 0, h \to 0, K \to \infty} q(\Delta, h, K) = 0,$$

such that for all $j = 0, \dots, M$, the following inequality holds:

$$\| z_j - y_j[K] \|_Y \leqslant Cq(\Delta, h, K).$$

The method (18) is said to be stable, if the operator $\hat{S}_K$ is Lipschitz with the constant $L_{\hat{S}_K}$ such that

$$L_{\hat{S}_K} = L_{\hat{S}_K}(\Delta, h, K) \leqslant 1. \tag{19}$$

The error of approximation of the method (18) is defined in a similar way.

The following theorem is hold, which could be proved in a similar way.

Theorem 2. *Let the method (18) is stable, the approximation error is of the order* $q(\Delta, h, K)$, $\lim\limits_{\Delta \to 0, h \to 0, K \to \infty} q(\Delta, h, K) = 0$, *then the method converges with the order* $q(\Delta, h, K)$.

[1] Notice that iterative method (11) has form (17).

4 Embedding of the Difference Method in a General Nonlinear Scheme

We rewrite the system (8) (or Eq. (10)) in the form

$$F(y_{j+1}) = \omega(y_{j+1}) - \Delta A y_{j+1} - \Delta A y_j - \omega(y_j) - \Delta f_{j+1/2} = 0. \tag{20}$$

Then Newton's method (9) (or (11)) could be written in the form

$$y_{j+1}[k+1] = y_{j+1}[k] - F'^{-1}(y_{j+1}[k])F(y_{j+1}[k]),$$
$$F'(y_{j+1}[k]) = \omega'(y_{j+1}[k]) - \Delta A,\ k = 0, \dots, K-1. \tag{21}$$

Let us denote

$$\Psi(y) = y - F'^{-1}(y)F(y). \tag{22}$$

Method (18) could be rewritten in the form

$$y_{j+1}[K] = \tilde{S}(\Psi(\Psi(\dots\Psi(y_{j+1}[0])))) + \Delta\tilde{\Phi}(\Psi(\Psi(\dots\Psi(y_{j+1}[0])))). \tag{23}$$

Taking into account (12), method (23) could also be written in the form

$$y_{j+1}[K] = \tilde{S}(\Psi(\Psi(\dots\Psi(y_j[K])))) + \Delta\tilde{\Phi}(\Psi(\Psi(\dots\Psi(y_{j+1}[0])))), \tag{24}$$

that coincides with form (18) if one takes $\hat{S}_K(\cdot) = \tilde{S}(\Psi(\Psi(\dots\Psi(\cdot))))$ and $\hat{\Phi}_K(\cdot) = \tilde{\Phi}(\Psi(\Psi(\dots\Psi(\cdot))))$.

The embedding of the method in the general scheme is done. Now we need to check the stability condition, show that operator $\hat{\Phi}_K$ is Lipschitz one and find out the order of the residual.

5 Stability and Approximation Order

Let us verify that under certain conditions the operator Ψ in (22) is contractive, which implies the stability condition in the approximation scheme. As a norm in the space Y hereinafter we use the Euclidean one.

Let y and $y+\varepsilon$ be two vectors from a neighborhood D_r of radius r with center at the root of Eq. (20). Then,

$$\Psi(y+\varepsilon) - \Psi(y) = y + \varepsilon - F'^{-1}(y+\varepsilon)F(y+\varepsilon) - y + F'^{-1}(y)F(y)$$
$$= \varepsilon - F'^{-1}(y+\varepsilon)F(y+\varepsilon) + F'^{-1}(y+\varepsilon)F(y) - F'^{-1}(y+\varepsilon)F(y) + F'^{-1}(y)F(y)$$
$$= \left\{\varepsilon - F'^{-1}(y+\varepsilon)(F(y+\varepsilon) - F(y))\right\} + \left\{F'^{-1}(y)F(y) - F'^{-1}(y+\varepsilon)F(y)\right\}. \tag{25}$$

Let us require the norms of expressions in each curly brace in (25) are less than $\frac{q}{2}\|\varepsilon\|$, $q < 1$, then the operator $\Psi(y)$ is contractive.

The following representation holds

$$F(y+\varepsilon) - F(y) = F'(y+\varepsilon)\varepsilon - \frac{1}{2}F''(\vartheta_{11})\varepsilon^2,$$

where $F''(y)$ is a diagonal matrix with $\omega''(y^i)$ on the diagonal, vector ϑ_{11} lies on the line segment connecting the tips of the vectors y and $y+\varepsilon$, and ε^2 is a vector with coordinates $(\varepsilon^i)^2$.

Since ω is twice continuously differentiable on the domain of our interest, $\omega''(y^i)$ is uniformly bounded, and therefore we have

$$|\omega''(u)| \le 2C_1.$$

Then, according to (7) the following estimates hold

$$\|F'^{-1}(y)\| \le C_2, \ \ \|F'^{-1}(y+\varepsilon)\| \le C_2. \tag{26}$$

Combining these gives us the following inequality

$$\|\varepsilon - F'^{-1}(y+\varepsilon)(F(y+\varepsilon) - F(y))\| \le C_1 C_2 \|\varepsilon\|^2.$$

Let us require the condition holds

$$C_1 C_2 \|\varepsilon\| < \frac{q}{2}, \tag{27}$$

this can be obtained by reducing r. For example, since $\|\varepsilon\| \le 2r$, then condition (27) is true if

$$4C_1 C_2 r < q. \tag{28}$$

As a result we obtain

$$\|\varepsilon - F'^{-1}(y+\varepsilon)(F(y+\varepsilon) - F(y))\| \le \frac{q}{2}\|\varepsilon\|. \tag{29}$$

Now let us estimate the second curly brace in (25)

$$\begin{aligned} F'^{-1}(y)F(y) - F'^{-1}(y+\varepsilon)F(y) &= F'^{-1}(y)(F'(y+\varepsilon) - F'(y))F'^{-1}(y+\varepsilon)F(y) \\ &= F'^{-1}(y)F''(\vartheta_{12})F'^{-1}(y+\varepsilon)F(y)\varepsilon, \end{aligned} \tag{30}$$

where vector ϑ_{12} lies on the segment connecting the tips of the vectors y and $y+\varepsilon$.

Since ω is twice smooth, there exist such constant C_3 that the following estimate holds

$$\|F''(\vartheta_{12})\| \leq C_3. \tag{31}$$

Let us estimate the term $F(y)$. If we take values on the previous layer as the initial approximation of Newton method on the current layer $y = y_{j+1}[0] = y_j[K]$, then (20) implies

$$F(y) = -2\Delta A y_j[K] - \Delta f_{j+1/2}.$$

Hence, in view of the assumption that Ay and $f_j(y)$ are bounded,

$$\|F(y)\| \leq \Delta C_4. \tag{32}$$

Take (30)–(32) into account and make a restriction on a step

$$\Delta C_2^2 C_3 C_4 < \frac{q}{2}, \tag{33}$$

then

$$\|S^{-1}(y)F(y) - S^{-1}(y+\varepsilon)F(y)\| \leq \frac{q}{2}\|\varepsilon\|. \tag{34}$$

As a result from (25), (29) and (34) we deduce

$$\|\Psi(y+\varepsilon) - \Psi(y)\| \leq L\|\varepsilon\|, \quad L < 1. \tag{35}$$

Note that condition (28) can also be rewritten in the form of a restriction on the smallness of a step Δ, just as it was done in (33).

Thus, it was proved the following

Lemma 2. *If conditions (28), (33) are satisfied, then the operator $\Psi(y)$ of the form (22) is contractive.*

Under the mentioned in Lemma 2 conditions method (9) converges and the following estimate holds

$$\|y_{j+1}[K] - y_{j+1}\| \leq q^K \|y_{j+1}[0] - y_{j+1}\|. \tag{36}$$

This also implies the

Theorem 3. *If conditions (28), (33) are met, then the implicit approximation method (9), represented in the form (18) is stable in the sense of definition (19).*

Proof. Since the operator $\Psi(y)$ is contractive and $\tilde{S}$ in (17) is bounded, then (24) implies the stability condition (19) in the approximation scheme. □

Let us make a remark. Operator $\hat{\Phi}_K$ defined according to (18), (24) is Lipschitz.

Now we study the residual of method (9) which is presented in form (18). First, consider method (8).

The residual of method (8) is, by definition, a grid function

$$\psi_j^i = \frac{\omega(u(x_i,t_{j+1})) - \omega(u(x_i,t_j))}{\Delta} - Au(x_i,t_j) - Au(x_i,t_{j+1}) - f_{j+1/2}, \quad (37)$$

where $i = 1,\dots,N-1,\ j = 0,\dots,M-1$.

Lemma 3. *Let the exact solution $u(x,t)$ of the boundary-value problem (4)–(6) be twice continuously differentiable with respect to t, all its derivatives with respect to x up to order four be continuous and belong to L^1, and let fractional derivative $\partial^\alpha u(x,t)/\partial x^\alpha$ be twice continuously differentiable with respect to t. Let also function $\omega(u)$ be twice continuously differentiable in a bounded domain containing the solution $u(x,t)$. Then there exists such a constant C_5, that*

$$|\psi_j^i| \le C_5(\Delta^2 + h),\ i = 1,\dots,N-1,\ j = 1,\dots,M-1. \quad (38)$$

Proof. Using the Taylor expansion with respect to t in the neighborhood of point $(x_i, t_{j+1/2})$, we have

$$\begin{aligned}\psi_j^i = \frac{1}{\Delta}\Bigg(&\omega(u(x_i,t_{j+1/2})) + \frac{\Delta}{2}\omega'(u(x_i,t_{j+1/2}))u_t'(x_i,t_{j+1/2})\\ &+\frac{\Delta^2}{8}\big(\omega''(u(x_i,t_{j+1/2}))u'(x_i,t_{j+1/2}) + \omega'(u(x_i,t_{j+1/2}))u''(x_i,t_{j+1/2})\big)\\ &-\Big[\omega(u(x_i,t_{j+1/2})) - \frac{\Delta}{2}\omega'(u(x_i,t_{j+1/2}))u_t'(x_i,t_{j+1/2})\\ &+\frac{\Delta^2}{8}\big(\omega''(u(x_i,t_{j+1/2}))u'(x_i,t_{j+1/2}) + \omega'(u(x_i,t_{j+1/2}))u''(x_i,t_{j+1/2})\big) + O(\Delta^3)\Big]\Bigg)\\ &-Au(x_i,t_j) - Au(x_i,t_{j+1}) - f(x_i,t_{j+1/2})\\ =& \frac{\partial\omega(u(x_i,t_{j+1/2}))}{\partial t} + O(\Delta^2) - Au(x_i,t_j) - Au(x_i,t_{j+1}) - f_{j+1/2}.\end{aligned}$$

According to [9]

$$Au(x_i,t_j) = \frac{\partial^\alpha u(x_i,t_j)}{\partial x^\alpha} + O(h).$$

Expanding these fractional derivatives in a series in powers of t in the neighbourhood of the point $(x_i, t_{j+1/2})$, we obtain (38). □

Let us now study how the solution of Eq. (8) changes when the parameters of the equation change. To do this along with an equation of the form (8) written in the form (20) we consider an equation of the form

$$\bar{F}(\bar{y}_{j+1}) = \omega(\bar{y}_{j+1}) - \Delta A\bar{y}_{j+1} - \Delta A\bar{y}_j - \omega(\bar{y}_j) - \Delta f_{j+1/2} = 0, \quad (39)$$

where $\|\bar{y}_j - y_j\| \le \delta$.

Lemma 4. *The following estimation holds*

$$\|\bar{y}_{j+1} - y_{j+1}\| \leq C_7\delta. \tag{40}$$

Proof. Subtract Eq. (20) from Eq. (39) and get

$$\omega(\bar{y}_{j+1}) - \Delta A\bar{y}_{j+1} - \Delta A\bar{y}_j - \omega(y_{j+1}) - \Delta Ay_{j+1} - \Delta Ay_j = \omega(\bar{y}_j) - \omega(y_j).$$

For the i-th coordinate this equation looks like

$$\omega(\bar{y}^i_{j+1}) - \Delta A\bar{y}^i_{j+1} - \Delta A\bar{y}^i_j - \omega(y^i_{j+1}) - \Delta Ay^i_{j+1} - \Delta Ay^i_j = \omega(\bar{y}^i_j) - \omega(y^i_j).$$

We use the finite-increments formula:

$$(\omega'(\theta) - \Delta A)(\bar{y}^i_{j+1} - y^i_{j+1}) = (\omega'(\theta_1) - \Delta A)(\bar{y}^i_j - y^i_j).$$

Due to condition (7) and operator's A properties there is such a constant C_7 that inequality (40) holds. □

Theorem 4. *If conditions (7), (28) and (33) hold then implicit approximation method (9), represented in form (18), has a residual of order* $\Delta^2 + h + \lambda^{2^K}$, *where* $0 < \lambda < 1$.

Proof. Let us rewrite expression (37) in the following form

$$\omega(u(x_i, t_{j+1})) - \Delta Au(x_i, t_{j+1}) = \omega(u(x_i, t_j)) + \Delta Au(x_i, t_j) + \Delta f_{j+1/2} + \Delta\hat{\psi}^i_j,$$

Due to Lemmas 3 and 4 the following estimate holds

$$|u(x_i, t_{j+1}) - u^i_{j+1}| \leq C_6C_7\Delta(\Delta^2 + h).$$

According to (36) this implies

$$|u(x_i, t_{j+1}[K]) - u^i_{j+1}[K]| \leq C_6C_7\Delta(\Delta^2 + h + \lambda^{2^K}),\ 0 < \lambda < 1,$$

and this implies the conclusion of the theorem. □

Theorems 2, 3 and 4 imply the theorem about convergence.

Theorem 5. *The implicit approximation method (9) written in the form (18) or (24) converges and has the order* $\Delta^2 + h + \lambda^{2^K},\ 0 < \lambda < 1$.

6 Numerical Examples

Let us consider two concrete examples. Namely, in Eq. (4) we take $\omega(u) = exp(u)$ in the first example and $\omega(u) = u + u^3$ in the second one.

Example 1. Let us consider the initial boundary value problem

$$\frac{\partial e^{u}}{\partial t} = \frac{\partial^{1.5} u}{\partial x^{1.5}} - e^{x^2 \cos t}\, x^2 \,\sin t - \frac{4}{\sqrt{\pi}} \sqrt{x}\, \cos t \tag{41}$$

on the domain $x \in (0, 1),\ t \in (0, 4\pi)$. Initial and boundary conditions are defined as follow

$$u(x, 0) = x^2, \quad 0 \le x \le 1,$$

$$u(0, t) = 0,\ u(1, t) = \cos t, \quad 0 \le t \le 4\pi.$$

Problem (41) has an exact solution $u(x, t) = x^2\, \cos t$.

Example 2. Let us consider the initial boundary value problem

$$\frac{\partial (u + u^3)}{\partial t} = \frac{\partial^{1.5} u}{\partial x^{1.5}} - (x^2 + 3x^6 \cos^2 t)\, \sin t - \frac{4}{\sqrt{\pi}} \sqrt{x}\, \cos t \tag{42}$$

on the domain $x \in (0, 1),\ t \in (0, 4\pi)$. Initial and boundary conditions are defined as follow

$$u(x, 0) = x^2, \quad 0 \le x \le 1,$$

$$u(0, t) = 0,\ u(1, t) = \cos t, \quad 0 \le t \le 4\pi.$$

Problem (41) has an exact solution $u(x, t) = x^2\, \cos t$.

In both examples the accuracy of Newton method was chosen to be $\epsilon = 10^{-5}$. The algorithm was implemented using Python 3.7, all computations were performed in a double precision.

Results of numerical Examples 1 and 2 are presented in Table 1. The third and fifth columns show the maximum of absolute difference between the exact and numerical solutions $\mathbf{diff}_{\Delta,h} = \max\limits_{i,j} |u^i_j - u(x_i, t_j)|,\ i = 0, \ldots, N,\ j = 0, \ldots, M,$ where N and M are the number of segments in space and time. The fourth and sixth columns show the ratio of the error reduction as the space grid refined.

In the series of experiments with $\Delta = \pi/40$ the error related to the time discretization is small in comparison with the error related to the coordinate discretization; the analysis of the error behavior reveals the first convergence with respect to space variables, i.e., when the step becomes half as much, the error becomes almost two times less as well.

The analysis of the data in the table shows that only the consistent decrease of steps yields the decrease of error. Indeed, in the series of experiments with $\Delta = \pi/10$ the halving of h does not cause the corresponding decrease of error, because the total error is mostly induced by the time discretization.

Table 1 Table of absolute errors and error rates. Numerical results of two examples are reported

Δ	h	$\omega(u) = exp(u)$ $\mathbf{diff}_{\Delta,h}$	Error rate	$\omega(u) = u + u^3$ $\mathbf{diff}_{\Delta,h}$	Error rate
$\pi/10$	1×2^{-2}	2.0207×10^{-2}	-	1.9947×10^{-2}	-
	1×2^{-3}	8.7812×10^{-3}	2.3012	8.8478×10^{-3}	2.2545
	1×2^{-4}	3.1322×10^{-3}	2.8035	3.1593×10^{-3}	2.8005
	1×2^{-5}	2.8955×10^{-3}	1.0817	3.2744×10^{-3}	0.9647
	1×2^{-6}	4.3512×10^{-3}	0.6654	4.6792×10^{-3}	0.6998
$\pi/20$	1×2^{-2}	2.3837×10^{-2}	-	2.3213×10^{-2}	-
	1×2^{-3}	1.2582×10^{-2}	1.8946	1.2282×10^{-2}	1.8901
	1×2^{-4}	6.0080×10^{-3}	2.0942	5.8997×10^{-3}	2.0818
	1×2^{-5}	2.4774×10^{-3}	2.4251	2.5062×10^{-3}	2.3540
	1×2^{-6}	7.8381×10^{-4}	3.1607	8.1704×10^{-4}	3.0675
$\pi/40$	1×2^{-2}	2.4683×10^{-2}	-	2.4027×10^{-2}	-
	1×2^{-3}	1.3501×10^{-2}	1.8282	1.3149×10^{-2}	1.8272
	1×2^{-4}	6.9687×10^{-3}	1.9374	6.7958×10^{-3}	1.9349
	1×2^{-5}	3.4219×10^{-3}	2.0365	3.3455×10^{-3}	2.0314
	1×2^{-6}	1.5745×10^{-3}	2.1732	1.5529×10^{-3}	2.1543

By Theorem 3 the proposed difference scheme is stable with any ratio of steps; however, due to the ill-posedness of the numerical differentiation, the decrease of h makes the approximations of $\partial^\alpha u/\partial^\alpha x$ in (4) more sensitive to the computer rounding error, which leads to the increase of the error. The decrease of Δ consistent with h is a peculiar regularizer which prevents errors from growing and accumulating. Experiments with $\Delta = \pi/10$ illustrate this fact.

Acknowledgements This work was supported by RFBR Grant 19-01-00019. The work was also supported by Act 211 Government of the Russian Federation, contract 02.A03.21.0006.

References

1. Gorbova, T.V., Pimenov, V.G., Solodushkin, S.I.: Difference schemes for the nonlinear equations in partial derivatives with heredity. In: Dimov, I., Farago, I., Vulkov, L. (eds.) Finite Difference Methods. Theory and Applications. FDM 2018. Lecture Notes in Computer Science, vol. 11386, pp. 258–265. Springer, Cham (2019)
2. Macias-Diaz, J., Hendy, A., De Staelen, R.: A pseudo energy-invariant method for relativistic wave equations with Riesz space-fractional derivatives. Comput. Phys. Commun. (2018). https://doi.org/10.1016/j.cpc.2017.11.008
3. Meerschaert, M.M., Tadjeran, C.: Finite difference approximations for two sided space fractional partial differential equations. Appl. Numer. Math. **65**, 80–90 (2006)
4. Pimenov, V., Hendy, A.: An implicit numerical method for the solution of the fractional advectiondiffusion equation with delay. Trudy Instituta Matematiki i Mekhaniki UrO RAN (2016). https://doi.org/10.1007/s001090000086

5. Pimenov, V., Hendy, A.: A fractional analog of Crank-Nicholson method for the two sided space fractional partial equation with functional delay. Ural Math. J. (2016). https://doi.org/10.15826/umj.2016.1.005
6. Pimenov, V.G.: General linear methods for the numerical solution of functional-differential equations. Differ. Equ. **37**(1), 116–127 (2001)
7. Podlubny, I.: Fractional Differential Equations: An Introduction to Fractional Derivatives, Fractional Differential Equations, to Methods of Their Solution and Some of Their Applications. Academic Press, Cambridge (1998)
8. Srivastava, V.K., Kumar, S., et al.: Two-dimensional time fractional-order biological population model and its analytical solution. Egypt J. Basic Appl. Sci. **1**, 71–76 (2014)
9. Tadjeran, C., Meerschaert, M.M., Scheffler, H.P.: A second-order accurate numerical approximation for the fractional diffusion equation. J. Comput. Phys. **213**, 205–214 (2006)
10. Samarskii, A.A.: The Theory of Difference Schemes. Taylor & Francis Inc, Milton Park (2001)

On the Behavior of Solutions with Positive Initial Data to Higher-Order Differential Equations with General Power-Law Nonlinearity

Tatiana Korchemkina

Abstract Higher-order differential equation with general power-law nonlinearity are considered. In particular, solutions with positive initial data are studied depending on the values of nonlinearity exponents. It is proven that if the sum of nonlinearity exponents is greater than one, then any considered solution has a finite right domain boundary. In the case of a constant potential solutions with power-law behavior are found.

1 Introduction

Consider solutions with positive initial data to higher order differential equation with general power-law nonlinearity

$$y^{(n)} = p\left(x,\ y,\ y',\ \ldots,\ y^{(n-1)}\right) |y|^{k_0} \left|y'\right|^{k_1} \ldots \left|y^{(n-1)}\right|^{k_{n-1}} \operatorname{sgn}\left(y\, y'\ \ldots y^{(n-1)}\right), \tag{1}$$

with $n \geq 2$, positive real nonlinearity exponents $k_0,\ k_1,\ \ldots k_{n-1}$ and positive continuous in x and Lipschitz continuous in $u_0, u_1, \ldots, u_{n-1}$ function $p(u_0, u_1, \ldots, u_{n-1})$ satisfying the inequalities

$$0 < m \leq p\left(x,\ y,\ y', \ldots,\ y^{(n-1)}\right) \leq M < +\infty. \tag{2}$$

Qualitative behavior and asymptotic estimates of positive increasing solutions for higher order differential equation $y^{(n)} = f\left(x, y, y', \ldots, y^{(n-1)}\right)$ with

$$(-1)^m f(x, u_0, u_1, \ldots, u_{n-1}) \geq g(x)|y^{(j)}|^{k_j}, \quad k_j > 1$$

were obtained by I.T. Kiguradze and T.A. Chanturia in [1]. Questions of qualitative and asymptotic behavior of solution to higher order Emden–Fowler differential equations ($k_1 = \ldots = k_{n-1} = 0$) were studied by I.V. Astashova in [2–5].

In the case $n = 2$ the results on qualitative behavior of solutions can be found in [6], and asymptotic behavior is studied in [7]. In this paper several results are generalized for higher order differential equations with general power-law nonlinearity.

T. Korchemkina (✉)
Lomonosov Moscow State University, Moscow, Russian Federation
e-mail: krtaalex@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_51

2 Qualitative Behavior of Solutions

Consider qualitative behavior of solutions with positive initial data. First of all, let us prove that $(n-1)$-th derivative of such solutions tends to infinity near their right domain boundaries.

Theorem 1. *Suppose that the function $p(u_0, u_1, \dots, u_{n-1})$ is continuous in x, Lipschitz continuous in $u_0, u_1, \dots, u_{n-1}$, and satisfies inequalities (2). Then for any maximally extended solution $y(x)$ to Eq. (1), satisfying the conditions $y(x_0) > 0$, $y'(x_0) > 0$, ..., $y^{(n-1)}(x_0) > 0$ at some point x_0, it holds that $y^{(n-1)} \to +\infty$ as $x \to x^*$, where $x^* \le +\infty$ is the right domain boundary of $y(x)$.*

Proof. Let us notice that

$$y^{(n)}(x) \ge m\ (y(x_0))^{k_0}\ \left(y'(x_0)\right)^{k_1} \dots \left(y^{(n-1)}(x_0)\right)^{k_{n-1}},$$

and then

$$y^{(n-1)}(x) - y^{(n-1)}(x_0) \ge m\ (y(x_0))^{k_0}\ \left(y'(x_0)\right)^{k_1} \dots \left(y^{(n-1)}(x_0)\right)^{k_{n-1}}\ (x - x_0).$$

Thus, in the case $x^* = +\infty$ derivative $y^{(n-1)}(x)$ is also unbounded as $x \to x^*$.

Consider now the case $x^* < +\infty$. Let us prove the statement of the theorem by contradiction: suppose $y^{(n-1)}(x) \le D_{n-1} < +\infty$ for $x \in [x_0, x^*)$. Then on this interval the following inequalities also hold:

$$y^{(n-2)}(x) \le D_{n-1}(x - x_0) + y^{(n-2)}(x_0) \le D_{n-1}(x^* - x_0) + y^{(n-2)}(x_0) = D_{n-2} < +\infty,$$

$$y^{(n-3)}(x) \le D_{n-2}(x - x_0) + y^{(n-3)}(x_0) \le D_{n-2}(x^* - x_0) + y^{(n-3)}(x_0) = D_{n-3} < +\infty,$$

and similarly we obtain that $y^{(j)}(x) \le D_j < +\infty$ for $j = 1, 2, \dots, n-1$, and also $y(x) \le D_0 < +\infty$. It means that the solution $y(x)$ and its derivatives have finite limits as $x \to x^* - 0$, which implies that the solution can be extended to the right of x^*, which leads to a contradiction.

Thus, $y^{(n-1)}(x) \to +\infty$ as $x \to x^*$, and the theorem is proven.

□

Now let us study whether the right domain boundary of a solution with positive initial data is finite or infinite. Denote

$$K = \sum_{i=0}^{n-1} k_i, \quad \varkappa = \sum_{i=1}^{n-1} i\, k_{n-1-i}.$$

Theorem 2. *Suppose $n \ge 2$, $K > 1$, the function $p(u_0, u_1, \dots, u_{n-1})$ is continuous in x, Lipschitz continuous in $u_0, u_1, \dots, u_{n-1}$, and satisfies inequality*

$$p\left(x,\ y,\ y',\ \dots,\ y^{(n-1)}\right) \ge m > 0.$$

Then there exists a constant $\xi = \xi(n, m, k_0, \dots, k_{n-1})$ *such that any maximally extended solution* $y(x)$ *to* (1)*, satisfying the conditions* $y(x_0) > 0$*,* $y'(x_0) > 0, \dots,$ $y^{(n-2)}(x_0) > 0$*,* $y^{(n-1)}(x_0) = y_{n-1} > 0$*, at some point* x_0 *has a finite right domain boundary* $x^* > x_0$ *and the following estimate holds:*

$$x^* - x_0 < \xi\, y_{n-1}^{-\frac{K-1}{\varkappa+1}}.$$

Proof. As it was shown in the previous theorem, $y^{(n-1)} \to +\infty$ as $x \to x^*$. Consider a sequence of points x_i, $i = 0, 1, \dots,$ such that $y^{(n-1)}(x_i) = 2\, y^{(n-1)}(x_{i-1}) = 2^i\, y_{n-1}$.

Then for $x \in [x_i, x_{i+1}]$ we have

$$y^{(n-1)} \ge 2^i\, y_{n-1},$$

hence

$$y^{(n-2)}(x) > y^{(n-2)}(x) - y^{(n-2)}(x_i) \ge 2^i\, y_{n-1}(x - x_i),$$

$$y^{(n-3)}(x) > y^{(n-3)}(x) - y^{(n-3)}(x_i) > 2^i\, y_{n-1} \frac{(x - x_i)^2}{2},$$

and by further integrating we obtain

$$y^{(j)}(x) > y^{(j)}(x) - y^{(j)}(x_i) \ge 2^i\, y_{n-1} \frac{(x - x_i)^{n-1-j}}{(n-1-j)!}, \quad j = n-2, \dots, 1,$$

and, finally,

$$y(x) > y(x) - y(x_i) > 2^i\, y_{n-1} \frac{(x - x_i)^{n-1}}{(n-1)!}.$$

Then, according to Eq. (1), for $y^{(n)}$ we derive

$$y^{(n)} > m \left| 2^i\, y_{n-1} \frac{(x - x_i)^{n-1}}{(n-1)!} \right|^{k_0} \dots \left| 2^i\, y_{n-1}(x - x_i) \right|^{k_{n-2}} \left| 2^i\, y_{n-1} \right|^{k_{n-1}}$$

$$y^{(n)} > C_0\, 2^{i\,K}\, y_{n-1}^K (x - x_i)^{\varkappa},$$

where C_0 is a constant depending only on m, n and $k_0, k_1, \dots, k_{n-1}$.

By integrating the above inequality on $[x_i, x_{i+1}]$ we obtain

$$2^i y_{n-1} > y^{(n-1)}(x_{i+1}) - y^{(n-1)}(x_i) > C_0\, (2^i y_{n-1})^K\, (x_{i+1} - x_i)^{\varkappa},$$

$$(x_{i+1} - x_i)^{\varkappa} < C_0^{-1} (2^i y_{n-1})^{-(K-1)},$$

and then

$$x_{i+1} - x_i < C_1 (2^i y_{n-1})^{-\frac{K-1}{\varkappa}}$$

with constant $C_1 = C_0^{-\frac{1}{\varkappa}}$ depending only on m, n and $k_0, k_1, \ldots, k_{n-1}$.

Now, summarizing the obtained inequalities for $i = 0, 1, \ldots$, we have

$$\sum_{i=0}^{+\infty} (x_{i+1} - x_i) < C_1 \, y_{n-1}^{-\frac{K-1}{\varkappa}} \sum_{i=0}^{+\infty} 2^{-i \frac{K-1}{\varkappa}}.$$

Since $K > 1$, the series in the right part of the above inequality converges, thus

$$x^* - x_0 = \lim_{i \to +\infty} x_i - x_0 = \sum_{i=0}^{+\infty} (x_{i+1} - x_i) < \xi \, y_{n-1}^{-\frac{K-1}{\varkappa}},$$

where $\xi = C_1 \sum\limits_{i=0}^{+\infty} 2^{-i \frac{K-1}{\varkappa}}$ is a constant depending only on $m, n, k_0, k_1, \ldots, k_{n-1}$, and the theorem is proven.

□

Remark 1. Theorem 2 generalizes the result obtained in [6] for $n = 2$.

3 On the Power-Law Solutions in the Case of a Constant Potential

Let us find a solution in the form $y = C(x^* - x)^{-\alpha}$ to Eq. (1) with constant potential $p\left(x, y, y', \ldots, y^{(n-1)}\right) \equiv (-1)^{n-1} p_0$.

Denote $\overline{\varkappa} = \sum\limits_{i=1}^{n-1} i \, k_i$.

Theorem 3. *Let $n \geq 2$, $p_0 > 0$ and $K > 1$. Then equation*

$$y^{(n)} = (-1)^{n-1} p_0 \, |y|^{k_0} \left|y'\right|^{k_1} \ldots \left|y^{(n-1)}\right|^{k_{n-1}} \operatorname{sgn}\left(y \, y' \ldots y^{(n-1)}\right) \tag{3}$$

has a solution $y = C(x^* - x)^{-\alpha}$, where $x^* < \infty$ is the right domain boundary,

$$C = \left(\frac{\prod\limits_{i=0}^{n-1} |\alpha + i|^{1 - \sum\limits_{i+1}^{n-1} k_i}}{p_0} \right)^{\frac{1}{K-1}}, \quad \alpha = \frac{n - \overline{\varkappa}}{K - 1}.$$

Proof. Let us put $y = C(x^* - x)^{-\alpha}$ into Eq. (3). According to the equation, we have

$$C(x^* - x)^{-\alpha-n} \prod_{i=0}^{n-1} (\alpha + i) = (-1)^{n-1} p_0 \operatorname{sgn}\big((-1)^{n-1} \alpha\,(\alpha + 1) \dots (\alpha + n - 2)\big) \cdot$$

$$\left|C(x^* - x)^{-\alpha}\right|^{k_0} \prod_{i=1}^{n-1} \left| C(x^* - x)^{-\alpha-i} \prod_{j=0}^{i-1} (\alpha + j) \right|^{k_i},$$

and due to that fact that $K > 1$,

$$\alpha + n - 1 = \frac{n - \overline{\varkappa}}{K - 1} + n - 1 =$$

$$= \frac{n - \sum\limits_{i=1}^{n-1} i k_i + (n-1)\left(\sum\limits_{i=0}^{n-1} k_i - 1\right)}{K - 1} = \frac{1 + \sum\limits_{i=0}^{n-1} (n - 1 - i)\, k_i}{K - 1} > 0,$$

which means that $\operatorname{sgn}(\alpha + n - 1) > 0$, and so

$$C(x^* - x)^{-\alpha-n} \prod_{i=0}^{n-1} |\alpha + i| = p_0\, C^K (x^* - x)^{-\sum\limits_{i=0}^{n-1} k_i(\alpha+i)} \prod_{i=1}^{n-1} \prod_{j=0}^{i-1} |\alpha + i|^{k_j},$$

which implies

$$(x^* - x)^{-\alpha-n+\sum\limits_{i=0}^{n-1} k_i(\alpha+i)} = p_0\, C^{K-1} \alpha^{-1} \prod_{i=1}^{n-1} \left(|\alpha + i|^{-1} \prod_{j=0}^{i-1} |\alpha + j|^{k_j} \right).$$

Since $x^* - x$ is a variable value, the equality is possible only when both left and right parts are equal to 1. Thus,

$$(x^* - x)^{-\alpha-n+\sum\limits_{i=0}^{n-1} k_i(\alpha+i)} = 1, \quad p_0\, C^{K-1} \alpha^{-1} \prod_{i=1}^{n-1} \left(|\alpha + i|^{-1} \prod_{j=0}^{i-1} |\alpha + j|^{k_j} \right) = 1, \tag{4}$$

and we derive the following equation for α:

$$\alpha + n = \sum_{i=0}^{n-1} k_i (\alpha + i),$$

hence

$$n - \overline{\varkappa} = \alpha\,(K - 1),$$

and

$$\alpha = \frac{n - \overline{\varkappa}}{K - 1} = \frac{n - k_1 - 2\,k_2 - \ldots - (n-1)\,k_{n-1}}{k_0 + k_1 + \ldots + k_{n-1} - 1}.$$

Now calculate the constant C from (4):

$$p_0\, C^{K-1} \alpha^{-1} \prod_{i=1}^{n-1} \left(|\alpha + i|^{-1} \prod_{j=0}^{i-1} |\alpha + j|^{k_j} \right) = 1,$$

$$C^{K-1} = \left(p_0\, \alpha^{-1} \prod_{i=1}^{n-1} \left(|\alpha + i|^{-1} \prod_{j=0}^{i-1} |\alpha + j|^{k_j} \right) \right)^{-1},$$

$$C^{K-1} = \left(\prod_{i=0}^{n-1} |\alpha + i| \right) \left(p_0 \prod_{i=0}^{n-1} |\alpha + i|^{\sum\limits_{i+1}^{n-1} k_i} \right)^{-1},$$

and, finally,

$$C = \left(\frac{\prod\limits_{i=0}^{n-1} |\alpha + i|^{1 - \sum\limits_{i+1}^{n-1} k_i}}{p_0} \right)^{\frac{1}{K-1}},$$

or

$$C = \left(\frac{|\alpha|^{1-k_1-\ldots-k_{n-1}} |\alpha + 1|^{1-k_2-\ldots-k_{n-1}} \ldots |\alpha + n - 2|^{1-k_{n-1}} |\alpha + n - 1|}{p_0} \right)^{\frac{1}{K-1}}.$$

□

Remark 2. Note that the equation

$$y^{(n)} = p_0 |y|^k \operatorname{sgn} y, \quad n \geq 2,\ k > 1,\ p_0 > 0$$

for any $x^* \in \mathbb{R}$ has the solution $y = C(x^* - x)^{-\alpha}$ with

$$\alpha = \frac{n}{k-1}, \quad C = \left(\frac{\alpha\,(\alpha + 1) \ldots (\alpha + n - 2)\,(\alpha + n - 1)}{p_0} \right)^{\frac{1}{k-1}},$$

which corresponds to the result obtained in theorem 3 with $k_1 = \ldots = k_{n-1} = 0$ (see [2], **5.1**). The existence of solutions to equation (1) equivalent to $C\,(x^* - x)^{-\alpha}$ as $x \to x^* - 0$ in general case is an open problem. For $n = 2$ this problem was solved in [7], and for $n \geq 3$, $k_1 = \ldots = k_{n-1} = 0$ it was solved in [2], Chap. 5 and [3, 5].

Acknowledgements The reported study was funded by RFBR, project number 19-31-90168.

References

1. Kiguradze, I.T., Chanturia, T.A.: Asymptotic Properties of Solutions of Nonautonomous Ordinary Differential Equations. Kluwer Academic Publishers, Dordrecht (1993)
2. Astashova, I.: Qualitative properties of solutions to quasilinear ordinary differential equations. In: Astashova, I.V. (ed.) Qualitative Properties of Solutions to Differential Equations and Related Topics of Spectral Analysis, Scientific edn., pp. 22–290. UNITY-DANA Publ, Moscow (2012). (in Russian)
3. Astashova, I.: On Kiguradze's problem on power-law asymptotic behavior of blow-up solutions to Emden-Fowler type differential equations. Georgian Math. J. **24**(2), 185–191 (2017)
4. Astashova, I.: On qualitative properties and asymptotic behavior of solutions to higher-order nonlinear differential equations. WSEAS Trans. Math. **5**(16), 39–47 (2017)
5. Astashova, I.V.: Asymptotic behavior of singular solutions of Emden-Fowler type equations. Differ. Equ. **55**(5), 581–590 (2019)
6. Korchemkina, T.: On the behavior of solutions to second-order differential equation with general power-law nonlinearity. Memoirs Differ. Equ. Math. Phys. **73**, 101–111 (2018)
7. Korchemkina, T.: On the asymptotic behavior of unbounded solutions to second order differential equations with general power-law nonlinearity. In: Proceedings os I.G. Petrovskii seminar, vol. 32, pp. 239–256 (2019) (in Russian)

Intuitionistic Fuzzy Stability of an Finite Dimensional Cubic Functional Equation

Sandra Pinelas, V. Govindan, K. Tamilvanan, and S. Baskaran

Abstract In the current work, the intuitionistic fuzzy version of Hypers-Ulam stability for a k-dimensional cubic functional equation

$$\sum_{j=1}^{k} f\left[\sum_{i(\neq j)=1}^{k} n^i x_i - n^j x_j\right] + (6-k) f\left(\sum_{i=1}^{k} n^i x_i\right)$$
$$=4\left[\sum_{j=1}^{k}\sum_{i(<j)=1}^{k} f(n^i x_i + n^j x_j) - (k-2)\sum_{i=1}^{k} f(n^i x_i)\right]$$

by applying a direct and fixed point methods is investigated. This way shows that some fixed points of a suitable operator can be a cubic mapping.

Keywords Cubic functional equation · Intuitionistic fuzzy normed space · Hyers-Ulam stability

2000 Mathematics Subject Classification: Primary 39B52 · 39B72 · 39B82

S. Pinelas
Departmento de Ciencias Exatas e Engenharia, Academia Militar, Lisbon, Portugal
e-mail: sandra.pinelas@gmail.com

V. Govindan (✉)
Department of Mathematics, Sri Vidya Mandir Arts and Science College,
Uthangarai 636 902, Tamil Nadu, India
e-mail: govindoviya@gmail.com

K. Tamilvanan
Department of Mathematics, Government Arts College, Krishnagiri 635001, Tamil Nadu, India
e-mail: tamiltamilk7@gmail.com

S. Baskaran
Department of Mathematics, Sri Meenakshi GGHSS, Tirupattur 635601, Tamil Nadu, India
e-mail: sps.baskaran@gmail.com

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_52

1 Introduction

In [29], Ulam proposed the general Ulam stability problem: "When is it true that by slightly changing the hypotheses of a theorem one can still assert that the thesis of the theorem remains true or approximately true?" In [14], Hyers gave the first affirmative answer to the question of Ulam for additive functional equations on Banach spaces. On the other hand, Cădariu and Radu noticed that a fixed point alternative method is very important for the solution of the Ulam problem. In other words, they employed this fixed point method to the investigation of the Cauchy functional equation [10] and for the quadratic functional equation [9] (for more applications of this method, see [3, 4, 6–8, 11] and [31]).

In 1965, Zadeh [32] introduced the notion of fuzzy sets which is a powerful hand set for modeling uncertainty and vagueness in various problems arising in the field of science and engineering. After that, fuzzy theory has become very active area of research and a lot of developments have been made in the theory of fuzzy sets to find the fuzzy analogues of the classical set theory. In fact, a large number of research papers have appeared by using the concept of fuzzy set and numbers and also fuzzification of many classical theories has been made. The concept of intuitionistic fuzzy normed spaces, initially has been introduced by Saadati and Park in [22]. Then, Saadati et al. have obtained a modified case of intuitionistic fuzzy normed spaces by improving the separation condition and strengthening some conditions in the definition of [2]. Many authors have considered the intuitionistic fuzzy normed linear spaces, and intuitionistic fuzzy 2-normed spaces (see [1, 2, 13, 16]). Also, the generalized Hyers-Ulam stability of different functional equations in intuitionistic fuzzy normed spaces has been studied by a number of the authors (see [5, 15, 20, 21, 25–28] and [30]).

In this paper, we consider the cubic functional equation of the form

$$\begin{aligned}\sum_{j=1}^{k} f\left[\sum_{i(\neq j)=1}^{k} n^i x_i - n^j x_j\right] + (6-k) f\left(\sum_{i=1}^{k} n^i x_i\right) \\ =4\left[\sum_{j=1}^{k}\sum_{i(<j)=1}^{k} f(n^i x_i + n^j x_j) - (k-2)\sum_{i=1}^{k} f(n^i x_i)\right]\end{aligned} \tag{1}$$

It is easy to check that the function $f(x) = ax^3$ is a solution of the functional Eq. (1) and also find the general solution of the cubic functional Eq. (1). In this paper, we study some stability results concerning the functional Eq. (1) in the setting intuitionistic fuzzy normed space.

2 Definitions and Notations

In this section, we firstly restate the usual terminology, notations and con- ventions of the theory of intuitionistic fuzzy normed space, as in [17], [19], [20], [21] and [23]. Then, we prove the generalized Ulam-Hyers stability of the Eq. (1) in intuitionistic fuzzy normed spaces, based on the fixed point Theorem.

Definition 2.1. An intuitionistic fuzzy set $A_{\zeta,\eta}$ in a universal set U is an object

$$A_{\zeta,\eta} = \{(\zeta_A(u), \eta_A(u))|u \in U\}$$

for all $u \in U$, $\zeta_A(u) \in [0, 1]$ and $\eta_A(u) \in [0, 1]$ are called the membership degree and the non-membership degree, respectively, of u in $A_{\zeta,\eta}$ and, furthermore, they satisfy $\zeta_A(u) + \eta_A(u) \leq 1$.

We denote its units by $0_{L^*} = (0, 1)$ and $1_{L^*} = (1, 0)$. Classically, a triangular norm $* = T$ on $[0, 1]$ is defined as an increasing, commutative, associative mapping $T : [0, 1]^2 \to [0, 1]$ satisfying $T(1, x) = 1 * x = x$ for all $x \in [0, 1]$. A triangular conorm $S = \Diamond$ is defined as an increasing, commutative, associative mapping $S : [0, 1]^2 \to [0, 1]$ satisfying $S(0, x) = 0\Diamond x = x$ for all $x \in [0, 1]$.

Using the lattice $(L^*, \leq_{L^*})$, these definitions can be straightforwardly extended.

Definition 2.2. A triangular norm (t-norm) on L^* is a mapping $T : (L^*)^2 \to L^*$ satisfying the following conditions:

(i) $(\forall x \in L^*)\ \ (T(x, 1_{L^*}) = x)$ (boundary condition);
(ii) $(\forall x, y \in (L^*)^2)(T(x, y) = T(y, x))$ (commutativity);
(iii) $(\forall x, y, z \in (L^*)^3)(T(x, T(y, z)) = T(T(x, y), z))$ (associativity);
(iv) $(\forall x, x', y, y' \in (L^*)^4)(x \leq_{L^*} x'$ and $y \leq_{L^*} y' \Rightarrow T(x, y) \leq_{L^*} T(x', y'))$ (monotonicity).

If $(L^*, \leq_{L^*}, T)$ is an Abelian topological monoid with unit 1_{L^*}, then L^* is said to be a continuous t-norm.

Definition 2.3. A continuous t-norms T on L^* is said to be continuous t-representable if there exist a continuous t-norm $*$ and a continuous t-conorm $\Diamond$ on $[0, 1]$ such that, for all $x = (x_1, x_2),\ y = (y_1, y_2) \in L^*$,

$$T(x, y) = (x_1 * y_1, x_2\Diamond y_2).$$

Definition 2.4. A negator on L^* is any decreasing mapping $N : L^* \to L^*$ satisfying $N : (0_{L^*}) = 1_{L^*}$ and $N(1_{L^*}) = 0_{L^*}$. If $N(N(x)) = x$ for all $x \in L^*$, then N is called an involutive negator. A negator on $[0, 1]$ is a decreasing mapping $N : [0, 1] \to [0, 1]$ satisfying $P_{\mu,\nu}(0) = 1$ and $P_{\mu,\nu}(1) = 0$. N_s denotes the standard negator on $[0, 1]$ defined by

$$N_s(x) = 1 - x, \quad \forall x \in [0, 1].$$

Definition 2.5. Let μ and ν be membership and nonmembership degree of an intuitionistic fuzzy set from $X \times (0, +\infty)$ to $[0, 1]$ such that $\mu_x(t) + \nu_x(t) \leq 1$ for all $x \in X$ and all $t > 0$. the triple $(X, P_{\mu,\nu}, T)$ is said to be an intuitionistic fuzzy normed space (briefly IFN-space) if X is a vector space, T is a continuous t-representable and $P_{\mu,\nu}$ is a mapping $X \times (0, +\infty) \to L^*$ satisfying the following conditions: for all $x, y \in X$ and $t, s > 0$,

(IFN1) $P_{\mu,\nu}(x, 0) = 0_{L^*}$;
(IFN2) $P_{\mu,\nu}(x, t) = 1_{L^*}$ if and only if $x = 0$;
(IFN3) $P_{\mu,\nu}(\alpha x, t) = P_{\mu,\nu}(x, \frac{t}{|\alpha|})$ for all $\alpha \neq 0$;
(IFN4) $P_{\mu,\nu}(x + y, t + s) \geq_{L^*} T(P_{\mu,\nu}(x, t), P_{\mu,\nu}(y, s))$.

In this case,$P_{\mu,\nu}$ is called an intuitionistic fuzzy norm. Here, $P_{\mu,\nu}(x, t) = (\mu_x(t), \nu_x(t))$.

Definition 2.6. A sequence $\{x_n\}$ in an IFN-space $(X, P_{\mu,\nu}, T)$ is called Cauchy sequence if, for any $\varepsilon > 0$ and $t > 0$, there exists $n_0 \in N$ such that

$$P_{\mu,\nu}(x_n - x_m, t) > L^* \ (N_s(\varepsilon), \varepsilon), \ \ \forall n, m \geq n_0,$$

where N_s is the standard negator.

Definition 2.7. The sequence $\{x_n\}$ is said to be convergent to a point $x \in X$ (denoted by $x_n \xrightarrow{P_{\mu,\nu}} x$) if

$$P_{\mu,\nu}(x_n - x, t) \to 1_{L^*} \ \ as \ \ n \to \infty$$

for every $t > 0$.

Definition 2.8. An IFN-space $(X, P_{\mu,\nu}, T)$ is said to be complete if every Cauchy sequence in X is convergent to a point $x \in X$.

Theorem (Banach Contraction Principle): Let (X, d) be a complete metric space and consider a mapping $T : X \to X$ which is strictly contractive mapping, that is

(A1) $d\left(T_x, T_y\right) \leq Ld\,(x, y)$ for some (Lipschitz constant) $L < 1$, then

(1) The mapping T has one and only fixed point $x^* = T\,(x^*)$;
(2) The fixed point for each given element x^* is globally attractive that is

(A2) $\lim\limits_{n\to\infty} T^n x = x^*$,for any starting point $x \in X$;

(1) One has the following estimation inequalities:

(A3) $d\,(T^n x, x^*) \leq \frac{1}{1-L} d\left(T^n x, T^{n+1} x\right)$,for all $n \geq 0, \ \ x \in X$.
(A4) $d\,(x, x^*) \leq \frac{1}{1-L} d\,(x, x^*)\,, \quad \forall\, x \in X$.

Theorem (The Alternative of fixed point): Suppose that for a complete generalized metric space (X, d) and a strictly contractive mapping $T : X \longrightarrow X$ with Lipschitz constant L. Then, for each given element $x \in X$ either

(B1) $d(T^n x, T^{n+1}x) = +\infty$, for all $n \geq 0$, or
(B2) There exists natural number n_0 such that

(i) $d(T^n x, T^{n+1}x) < \infty$ for all $n \geq n_0$;
(ii) The sequence $(T^n x)$ is convergent to a fixed point y^* of T;
(iii) y^* is the unique fixed point of T in the set $Y = \{y \in X; d(T^{n_0}x, y) < \infty\}$;
(iv) $d(y^*, y) \leq \frac{1}{1-L} d(y, Ty)$ for all $y \in Y$.

3 General Solution of the Functional Eq. (1)

In this section, we discuss the general solution of the functional Eq. (1).

Theorem 3.1. *If an odd mapping* $f : X \to Y$ *satisfies the functional equation*

$$f(2x + y) + f(2x - y) = 2f(x + y) + 2f(x - y) + 12f(x), \ \forall x, y \in X \quad (2)$$

if and only if $f : X \to Y$ *satisfies the functional equation*

$$\begin{aligned} &\sum_{j=1}^{k} f\left[\sum_{i(\neq j)=1}^{k} n^i x_i - n^j x_j\right] + (6-k) f\left(\sum_{i=1}^{k} n^i x_i\right) \\ &= 4\left[\sum_{j=1}^{k} \sum_{i(<j)=1}^{k} f(n^i x_i + n^j x_j) - (k-2) \sum_{i=1}^{k} f(n^i x_i)\right] \end{aligned} \quad (3)$$

forall $x_1, x_2, ...x_n \in X$

Proof. Let $f : X \to Y$ satisfies the functional Eq. (2). Setting (x, y) by $(0, 0)$ in (2), we get $f(0) = 0$. Replacing (x, y) by $(x, 0)$, (x, x) and $(x, 2x)$ respectively in (2), we obtain

$$f(2x) = 2^3 f(x), \quad f(3x) = 3^3 f(x) \ \ and \ \ f(4x) = 4^3 f(x) \quad (4)$$

for all $x \in X$. In general for any positive integer a, we have

$$f(ax) = a^3 f(x), \quad (5)$$

for all $x \in X$. It is easy to verify from (5), that

$$f(a^2 x) = a^6 f(x) \ \ and \ \ f(a^3 x) = a^9 f(x), \quad (6)$$

for all $x \in X$. Pluging (x, y) by $(nx_1 + n^2x_2, n^3x_3)$ in (2), we get

$$\begin{aligned} &f(2nx_1 + 2n^2x_2 + n^3x_3) + f(2nx_1 + 2n^2x_2 - n^3x_3) + 2f(-nx_1 - n^2x_2 - n^3x_3) \\ &\quad + 2f(-nx_1 - n^2x_2 + n^3x_3) = 12f(nx_1 + n^2x_2) \end{aligned} \tag{7}$$

for all $x_1, x_2, x_3 \in X$. Replacing (x, y) by $(n^2x_2 + n^3x_3, nx_1)$ in (2), we have

$$\begin{aligned} &f(nx_1 + 2n^2x_2 + n_3^x) + f(-nx_1 + 2n^2x_2 + 2n^3x_3) + 2f(-nx_1 - n^2x_2 - n^3x_3) \\ &\quad + 2f(nx_1 - n^2x_2 - n^3x_3) = 12f(n^2x_2 + n^3x_3) \end{aligned} \tag{8}$$

for all $x_1, x_2, x_3 \in X$. Switching (x, y) by $(nx_1 + n^3x_3, n^2x_2)$ in (2), we arrive

$$\begin{aligned} &f(2nx_1 + n^2x_2 + 2n^3x_3) + f(2nx_1 - n^2x_2 + 2n^3x_3) + 2f(-nx_1 - n^2x_2 - n^3x_3) \\ &\qquad + 2f(-nx_1 + n^2x_2 - n^3x_3) = 12f(nx_1 + n^3x_3) \end{aligned} \tag{9}$$

for all $x_1, x_2, x_3 \in X$. Adding (7), (8) and (9), We obtain

$$\begin{aligned} &12f(nx_1 + n^2x_2) + 12f(n^2x_2 + n^3x_3) + 12f(nx_1 + n^3x_3) = f(2nx_1 + 2n^2x_2 + n^3x_3) \\ &+ f(2nx_1 + 2n^2x_2 - n^3x_3) + 2f(-nx_1 - n^2x_2 - n^3x_3) + 2f(-nx_1 - n^2x_2 + n^3x_3) \\ &+ f(nx_1 + 2n^2x_2 + 2n^3x_3) + f(-nx_1 - 2n^2x_2 + 2n^3x_3) + 2f(-nx_1 - n^2x_2 - n^3x_3) \\ &+ 2f(nx_1 - n^2x_2 - n^3x_3) + f(2nx_1 + n^2x_2 + 2n^3x_3) + f(2nx_1 - n^2x_2 + 2n^3x_3) \\ &+ 2f(-nx_1 - n^2x_2 - n^3x_3) + 2f(-nx_1 + n^2x_2 - mmn^3x_3) \end{aligned} \tag{10}$$

for all $x_1, x_2, x_3 \in X$. Replacing (x, y) by $(nx_1, 2n^2x_2 + n^3x_3)$ in (2), we reach

$$\begin{aligned} &f(2nx_1 + 2n^2x_2 + n^3x_3) = f(-2nx_1 + 2n^2x_2 + n^3x_3) + 2f(nx_1 + 2n^2x_2 \\ &+ n^3x_3) + 2f(nx_1 - 2n^2x_2 - n^3x_3) + 12f(nx_1) \end{aligned} \tag{11}$$

for all $x_1, x_2, x_3 \in X$. Adding $f(2nx_1 + 2n^2x_2 - n^3x_3)$ on both sides of (11), we get

$$\begin{aligned} &f(2nx_1 + 2n^2x_2 + n^3x_3) + f(2nx_1 + 2n^2x_2 - n^3x_3) = 2f(nx_1 + 2n^2x_2 \\ &+ n^3x_3) + 2f(nx_1 - 2n^2x_2 - n^3x_3) + 12f(nx_1) + 2f(-2nx_1 + n^2x_2 + n^3x_3) \\ &+ 2f(2nx_1 + n^2x_2 - n^3x_3) + 12f(n^2x_2) \end{aligned} \tag{12}$$

for all $x_1, x_2, x_3 \in X$. Interchanging (x, y) by $(n^2x_2, nx_1 + 2n^3x_3)$ in (2), we have

$$\begin{aligned} &f(nx_1 + 2n^2x_2 + 2n^3x_3) = f(nx_1 + 2n^2x_2 + 2n^3x_3) + 2f(-nx_1 + n^2x_2 \\ &- 2n^3x_3) + 12f(n^2x_2) + f(nx_1 - 2n^2x_2 + 2n^3x_3) \end{aligned} \tag{13}$$

for all $x_1, x_2, x_3 \in X$. Adding $f(-nx_1 + 2n^2x_2 + 2n^3x_3)$ on both sides of (13), we arrive

$$\begin{aligned} f(nx_1+2n^2x_2+2n^3x_3)+f(-nx_1+2n^2x_2+2n^3x_3)&=2f(nx_1+n^2x_2\\ +2n^3x_3)+2f(-nx_1+n^2x_2-2n^3x_3)+12f(n^2x_2)+2f(nx_1-2n^2x_2\\ +n^3x_3)+2f(-nx_1+2n^2x_2+n^3x_3)+12f(n^3x_3) \end{aligned} \tag{14}$$

for all $x_1, x_2, x_3 \in X$. Pluging (x, y) by $(n^3x_3, 2nx_1+n^2x_2)$ in (2), we receive

$$\begin{aligned} f(2nx_1+n^2x_2+2n^3x_3)=2f(2nx_1+n^2x_2+n^3x_3)1+2f(-2nx_1-n^2x_2\\ +n^3x_3)+12f(n^3x_3)+f(2nx_1+n^2x_2-2n^3x_3) \end{aligned} \tag{15}$$

for all $x_1, x_2, x_3 \in X$. Adding $f(2nx_1-n^2x_2+2n^3x_3)$ on both sides of (15), we have

$$\begin{aligned} f(2nx_1+n^2x_2+2n^3x_3)+f(2nx_1-n^2x_2+2n^3x_3)=2f(2nx_1+n^2x_2+n^3x_3)\\ +2f(-2nx_1-n^2x_2+n^3x_3)+12f(n^3x_3)+2f(nx_1+n^2x_2-2n^3x_3)\\ +2f(nx_1-n^2x_2+2n^3x_3)+12f(nx_1) \end{aligned} \tag{16}$$

for all $x_1, x_2, x_3 \in X$. Using (12), (14) and (16) in (10), we arrive

$$\begin{aligned} 12f(nx_1+n^2x_2)+12f(n^2x_2+n^3x_3)+12f(nx_1+n^3x_3)=2f(nx_1+2n^2x_2+n^3x_3)\\ +2f(nx_1-2n^2x_2-n^3x_3)+12f(nx_1)+2f(-2nx_1+n^2x_2+n^3x_3)\\ +2f(2nx_1+n^2x_2-n^3x_3)+12f(n^2x_2)+2f(-nx_1-n^2x_2-n^3x_3)+2f(-nx_1-n^2x_2\\ +n^3x_3)+2f(nx_1+n^2x_2+2n^3x_3)+2f(-nx_1+n^2x_2-2n^3x_3)+12f(n^2x_2)\\ +2f(nx_1-2n^2x_2+n^3x_3)+2f(-nx_1+2n^2x_2+n^3x_3)+12f(n^3x_3)\\ +2f(-nx_1-n^2x_2-n^3x_3)+2f(nx_1-n^2x_2-n^3x_3)+2f(2nx_1+n^2x_2+n^3x_3)\\ +2f(-2nx_1-n^2x_2+n^3x_3)+12f(n^3x_3)+2f(nx_1+n^2x_2-2n^3x_3)\\ +2f(nx_1-n^2x_2+2n^3x_3)+12f(nx_1)+2f(-nx_1-n^2x_2-n^3x_3)\\ +2f(-nx_1+n^2x_2-n^3x_3) \end{aligned} \tag{17}$$

for all $x_1, x_2, x_3 \in X$. Replacing (x, y) by $(n^3x_3, 2nx_1+n^3x_3)$ in (2), we get

$$\begin{aligned} f(2nx_1+2n^2x_2+n^3x_3)=2f(2nx_1+n^2x_2+n^3x_3)+2f(-2nx_1+n^2x_2\\ -n^3x_3)+12f(n^2x_2)+f(2nx_1-2n^2x_2+n^3x_3) \end{aligned} \tag{18}$$

for all $x_1, x_2, x_3 \in X$. Adding $f(2nx_1+2n^2x_2-n^3x_3)$ on both sides of (18), we reach

$$\begin{aligned} f(2nx_1+2n^2x_2+n^3x_3)+f(2nx_1+2n^2x_2-n^3x_3)=2f(2nx_1+n^2x_2+n^3x_3)\\ +2f(-2nx_1-n^2x_2-n^3x_3)+12f(n^2x_2)+2f(nx_1-2n^2x_2+n^3x_3)\\ +2f(nx_1+2n^2x_2-n^3x_3)+12f(nx_1) \end{aligned} \tag{19}$$

for all $x_1, x_2, x_3 \in X$. Switching (x, y) by $(n^3x_3, nx_1 + 2n^2x_2)$ in (2), we obtain

$$\begin{aligned} f(nx_1 + 2n^2x_2 + 2n^3x_3) &= 2f(nx_1 + 2n^2x_2 + n^3x_3) + 2f(-nx_1 - 2n^2x_2 \\ &\quad +n^3x_3) + 12f(n^3x_3) + f(nx_1 + 2n^2x_2 - 2n^3x_3) \end{aligned} \tag{20}$$

for all $x_1, x_2, x_3 \in X$. Adding $f(-nx_1 + 2n^2x_2 + 2n^3x_3)$ on both sides of (20), we attain

$$\begin{aligned} f(nx_1 + 2n^2x_2 + 2n^3x_3) + f(-nx_1 + 2n^2x_2 + 2n^3x_3) &= 2f(2nx_1 + 2n^2x_2 + n^3x_3) \\ +2f(-nx_1 - 2n^2x_2 + n^3x_3) + 12f(n^3x_3) &+ 2f(nx_1 + n^2x_2 - 2n^3x_3) \\ +2f(-nx_1 + n^2x_2 + 2n^3x_3) &+ 12f(n^2x_2) \end{aligned} \tag{21}$$

for all $x_1, x_2, x_3 \in X$. Replacing (x, y) by $(nx_1, n^2x_2 + 2n^3x_3)$ in (2), we reach

$$\begin{aligned} f(2nx_1 + n^2x_2 + 2n^3x_3) &= 2f(nx_1 + n^2x_2 + 2n^3x_3) + 2f(nx_1 - n^2x_2 \\ -2n^3x_3) + 12f(nx_1) &+ f(-2nx_1 + n^2x_2 + 2n^3x_3) \end{aligned} \tag{22}$$

for all $x_1, x_2, x_3 \in X$. Adding $f(2nx_1 - n^2x_2 + 2n^3x_3)$ on both sides of (22), we arrive

$$\begin{aligned} f(2nx_1 + n^2x_2 + 2n^3x_3) + f(2nx_1 - n^2x_2 + 2n^3x_3 &= 2f(2nx_1 + n^2x_2 + n^3x_3)) \\ +2f(nx_1 - n^2x_2 - 2n^3x_3) + 12f(nx_1) &+ 2f(-2nx_1 + n^2x_2 + n^3x_3) \\ +2f(2nx_1 - n^2x_2 + n^3x_3) &+ 12f(n^3x_3) \end{aligned} \tag{23}$$

for all $x_1, x_2, x_3 \in X$. Using (19), (21) and (23) in (20), we have

$$\begin{aligned} &12f(nx_1 + n^2x_2) + 12f(n^2x_2 + n^3x_3) + 12f(nx_1 + n^3x_3) \\ &= 2f(nx_1 - 2n^2x_2 + n^3x_3) + 2f(nx_1 + 2n^2x_2 - n^3x_3) \\ &+12f(nx_1) + 2f(2nx_1 + n^2x_2 + n^3x_3) + 2f(-2nx_1 + n^2x_2 - n^3x_3) \\ &+12f(n^2x_2) + 2f(-nx_1 - n^2x_2 - n^3x_3) + 2f(-nx_1 - n^2x_2 \\ &+n^3x_3) + 2f(nx_1 + n^2x_2 - 2n^3x_3) + 2f(-nx_1 + n^2x_2 + 2n^3x_3) \\ &+12f(n^2x_2) + 2f(nx_1 + 2n^2x_2 + n^3x_3) + 2f(-nx_1 - 2n^2x_2 + n^3x_3) \\ &12f(n^3x_3) + 2f(-nx_1 - n^2x_2 - n^3x_3) + 2f(nx_1 - n^2x_2 - n^3x_3) \\ &+2f(-2nx_1 + n^2x_2 + n^3x_3) + 2f(2nx_1 - n^2x_2 + n^3x_3) + 12f(n^3x_3) \\ &+2f(nx_1 + n^2x_2 + 2n^3x_3) + 2f(nx_1 - n^2x_2 - 2n^3x_3) + 12f(nx_1) \\ &+2f(-nx_1 - n^2x_2 - n^3x_3) + 2f(-nx_1 + n^2x_2 - n^3x_3) \end{aligned} \tag{24}$$

for all $x_1, x_2, x_3 \in X$. Adding Eqs. (17) , (24) and continuing this process upto n times attacks, we arrive

$$\begin{aligned}\sum_{j=1}^{k} f\left[\sum_{i(\neq j)=1}^{k} n^i x_i - n^j x_j\right] + (6-k) f\left(\sum_{i=1}^{k} n^i x_i\right) \\ = 4\left[\sum_{j=1}^{k}\sum_{i(<j)=1}^{k} f(n^i x_i + n^j x_j) - (k-2)\sum_{i=1}^{k} f(n^i x_i)\right]\end{aligned} \tag{25}$$

for all $x_i, x_j \in X$.Conversely, $f : X \to Y$ satisfies the functional Eq. (3) and replacing $(x_1, x_2, x_3, ...x_k)$ by $(x, 0, 0.....0)$, $(0, x, 0, 0.....0)$ *and* $(0, 0, x, 0, 0....0)$ respectively in (3), we obtain

$$f(nx) = n^3 f(x), \quad f(n^2x) = n^6 f(x), \quad f(n^3x) = n^9 f(x) \tag{26}$$

and so on, for all $x \in X$. One can easy to verify from (26) that

$$f\left(\frac{x}{n^i}\right) = \left(\frac{1}{n^i}\right)^3 f(x_i), \quad \forall x, y \in X. \tag{27}$$

Switching $(x_1, x_2, x_3, ...x_k)$ by $\left(\frac{x}{n}, \frac{x}{n^2}, \frac{y}{n^3}, 0, 0....0\right)$ in (3), we get

$$3f(2x+y) + f(2x-y) = 24f(x) - 6f(y) + 8f(x+y) \tag{28}$$

for all $x, y \in X$. Interchanging y by $-y$ in (28), we have

$$3f(2x-y) + f(2x+y) = 24f(x) + 6f(y) + 8f(x-y) \tag{29}$$

for all $x, y \in X$. Adding the Eqs. (28) and (29), we arrive our result. □

4 Stability Results for the Functional Eq. (1): Direct Method

In this section, we investigate the generalized Ulam-Hyers stability of the functional Eq. (1) in Intuitionistic fuzzy normed space via direct method.

Theorem 4.1. *Let* $\beta \in \{-1, 1\}$. *Let* X *be a linear space,* $(Z, P'_{\mu,\nu}, T)$ *be an IFN-space,* $\alpha : X^n \to Z$ *be a mapping with* $0 < \left(\frac{d}{2^3}\right)^\beta < 1$,

$$P'_{\mu,\nu}\left(\alpha(2^\beta x, 0, 0), r\right) \geq_{L^*} P'_{\mu,\nu}\left(d^\beta \alpha(x, 0, \cdots, 0), r\right) \tag{30}$$

for all $x \in X$ *and all* $r > 0$, *and*

$$\lim_{n\to\infty} P'_{\mu,\nu}\left(\alpha(2^{\beta n}x_1, 2^{\beta n}x_2, \cdots, 2^{\beta n}x_k), 2^{3\beta n}r\right) = 1_{L^*} \tag{31}$$

for all $x_1, x_2, \cdots, x_k \in X$ *and all* $r > 0$. $\left(Y, P'_{\mu,\nu}, T\right)$ *be an IFN-space. Suppose that a function* $f : X \to Y$ *satisfies the inequality*

$$P_{\mu,\nu}\left(Df(x_1, x_2, \cdots, x_k), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x_1, x_2, \cdots, x_k), r\right) \tag{32}$$

for all $r > 0$ *and all* $x_1, x_2, \cdots, x_k \in X$. *Then the limit*

$$P_{\mu,\nu}\left(C(x) - \frac{f(2^{\beta n}x)}{2^{3\beta n}}\right) \to 1_{L^*}, \quad as \quad n \to \infty, \quad r > 0 \tag{33}$$

exists for all $x \in X$ *and the mapping* $C : X \to Y$ *is a unique cubic mapping satisfying (1) and*

$$P_{\mu,\nu}\left(f(x) - C(x), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), 4r \mid n^3 - d \mid\right) \tag{34}$$

for all $x \in X$ *and all* $r > 0$.

Proof. First assume $\beta = 1$. Replacing $(x_1, x_2, \cdots, x_k)$ by $(x, 0, \cdots, 0)$ in (1), we get

$$P_{\mu,\nu}\left(4f(nx) - 4n^3 f(x), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), r\right) \tag{35}$$

for all $x \in X$ and all $r > 0$. Replacing x by $n^i x$ in (35) and using (IFN3), we have

$$P_{\mu,\nu}\left(\frac{f(n^{i+1}x)}{n^3} - f(n^i x), \frac{r}{4n^3}\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(n^i x, 0, \cdots, 0), r\right) \tag{36}$$

for all $x \in X$ and all $r > 0$. Using (30), (IFN3) in (36), we get

$$P_{\mu,\nu}\left(\frac{f(n^{i+1}x)}{n^3} - f(n^i), \frac{r}{4n^3}\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), \frac{r}{d^i}\right) \tag{37}$$

for all $x \in X$ and all $r > 0$. It is easy to verify from (37), that

$$P_{\mu,\nu}\left(\frac{f(n^{i+1}x)}{n^{3(i+1)}} - \frac{f(n^i)}{n^{3i}}, \frac{r}{4n^{3(i+1)}}\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), \frac{r}{d^i}\right) \tag{38}$$

holds for all $x \in X$ and all $r > 0$. Replacing r by $d^i r$ in (38), we obtain

$$P_{\mu,\nu}\left(\frac{f(n^{i+1}x)}{n^{3(i+1)}} - \frac{f(n^i)}{n^{3i}}, \frac{d^i r}{4n^{3(i+1)}}\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), r\right) \tag{39}$$

for all $x \in X$ and all $r > 0$. It is easy to see that

$$\frac{f(2^i x)}{2^{3i}} - f(x) = \sum_{l=0}^{i-1} \frac{f(n^{l+1}x)}{n^{3(l+1)}} - \frac{f(n^l x)}{n^{3l}} \tag{40}$$

for all $x \in X$. from Eqs. (39) and (40), we get

$$P_{\mu,\nu}\left(\frac{f(n^i x)}{n^{3i}} - f(x), \sum_{l=0}^{i-1} \frac{d^l r}{4n^{3(l+1)}}\right) \geq_{L^*} T_{l=0}^{i-1}\left(P^{'}_{\mu,\nu}\left(\frac{f(n^{l+1}x)}{n^{3(l+1)}} - \frac{f(n^l x)}{n^{3l}}, \frac{d^l r}{n^{3(l+1)}}\right)\right)$$

$$\geq_{L^*} T_{l=0}^{i-1}\left(P^{'}_{\mu,\nu}(\alpha(x,0,\cdots,0),r)\right)$$

$$\geq_{L^*} P^{'}_{\mu,\nu}(\alpha(x,0,\cdots,0),r) \tag{41}$$

for all $x \in X$ and all $r > 0$. Replacing x by $n^m x$ in (41) and using (30), we attain

$$P_{\mu,\nu}\left(\frac{f(n^{i+m}x)}{n^{3(i+m)}} - \frac{f(n^m x)}{n^{3m}}, \sum_{l=0}^{i-1} \frac{d^l r}{4n^{3(1+l+m)}}\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\alpha(x,0,\cdots,0), \frac{r}{d^m}\right) \tag{42}$$

for all $x \in X$ and all $r > 0$ and all $m, i \geq 0$. Replacing r by $d^m r$ in (42), we get

$$P_{\mu,\nu}\left(\frac{f(n^{i+m}x)}{n^{3(i+m)}} - \frac{f(n^m x)}{n^{3m}}, \sum_{l=m}^{i+m-1} \frac{d^l r}{4n^{3(1+l)}}\right) \geq_{L^*} P^{'}_{\mu,\nu}(\alpha(x,0,\cdots,0),r) \tag{43}$$

for all $x \in X$ and all $r > 0$ and all $m, i \geq 0$. Using (IFN3) in (43), we reach

$$P_{\mu,\nu}\left(\frac{f(n^{i+m}x)}{n^{3(i+m)}} - \frac{f(n^m x)}{n^{3m}}, r\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\alpha(x,0,\cdots,0), \frac{r}{\sum_{l=m}^{i+m-1} \frac{d^l}{4n^{3(1+l)}}}\right) \tag{44}$$

for all $x \in X$ and all $r > 0$ and all $m, i \geq 0$. Since $0 < d < n^3$ and $\sum_{l=0}^{i}\left(\frac{d}{n^3}\right)^l < \infty$. Thus $\{\frac{f(n^i x)}{n^{3i}}\}$ is a Cauchy sequence in $(Y, P_{\mu,\nu}, T)$. Since $(Y, P_{\mu,\nu}, T)$ is a complete IFN-space, this sequence converges to some point $C(x) \in Y$. So one can define the mapping $C : X \to Y$ by

$$P_{\mu,\nu}\left(C(x) - \frac{f(n^{\beta i}x)}{n^{3\beta i}}\right) \to 1_{L^*} \quad as \quad n \to \infty, \quad r > 0$$

for all $x \in X$. Letting $m = 0$ in (44), we get

$$P_{\mu,\nu}\left(\frac{f(n^i x)}{n^{3i}} - f(x), r\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\alpha(x,0,\cdots,0), \frac{r}{\sum_{l=0}^{i-1} \frac{d^l}{4n^{3(l+1)}}}\right) \tag{45}$$

for all $x \in X$ and all $r > 0$. Letting $n \to \infty$ in (45), we reach

$$P_{\mu,\nu}(f(x) - C(x), r) \geq_{L^*} P^{'}_{\mu,\nu}\left(\alpha(x,0,\cdots,0), 4r(n^3 - d)\right)$$

for all $x \in X$ and all $r > 0$. To prove C satisfies the (1), replacing $(x_1, x_2, \cdots, x_k)$ by $(n^i x_1, n^i x_2, \cdots, n^i x_k)$ in (32), respectively, we attain

$$P_{\mu,\nu}\left(\frac{1}{n^{3i}} Df(n^i x_1, n^i x_2, \cdots, n^i x_k), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(n^i x_1, n^i x_2, \cdots, n^i x_k), n^{3i} r\right) \tag{46}$$

for all $r > 0$ and all $x_1, x_2, \cdots, x_k \in X$. Hence C satisfies the cubic functional Eq. (1). In order to prove $C(x)$ is unique, we let $D(x)$ be another cubic functional equation satisfying (1) and (34). Hence,

$$P_{\mu,\nu}(C(x) - D(x), r) = P_{\mu,\nu}\left(\frac{C(n^i x)}{n^{3i}} - \frac{D(n^i x)}{n^{3i}}, r\right)$$

$$\geq_{L^*} T\{P_{\mu,\nu}\left(\frac{C(n^i x)}{n^{3i}} - \frac{f(n^i x)}{n^{3i}}, \frac{r}{2}\right), P_{\mu,\nu}\left(\frac{f(n^i x)}{n^{3i}} - \frac{D(n^i x)}{n^{3i}}, \frac{r}{2}\right), \frac{r}{2}\}$$

$$\geq_{L^*} P'_{\mu,\nu}\left(\alpha(n^i x, 0, \cdots, 0), \frac{4rn^{3i}(n^3 - d)}{2}\right)$$

$$\geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), \frac{2rn^{3i}(n^3 - d)}{d^i}\right)$$

for all $x \in X$ and all $r > 0$. Since

$$\lim_{n\to\infty} \frac{2rn^{3i}(n^3 - d)}{d^i} = \infty,$$

we obtain

$$\lim_{n\to\infty} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), \frac{2rn^{3i}(n^3 - d)}{d^i}\right) = 1_{L^*}.$$

Thus

$$P_{\mu,\nu}(C(x) - D(x), r) = 1_{L^*}$$

for all $x \in X$ and all $r > 0$, hence $C(x) = D(x)$. Therefore $C(x)$ is unique. For $\beta = -1$, we can prove the result by a similar method. This completes the proof of the theorem. □

The following corollary is an immediate consequence of Theorem 4.1, concerning the stability for the functional Eq. (1).

Corollary 4.1. *Suppose that the function $f : X \to Y$ satisfies the inequality*

$$P_{\mu,\nu}(Df(x_1,x_2,\cdots,x_k),r) \geq_{L^*} \begin{cases} P^{'}_{\mu,\nu}(\varphi, r) \\ P^{'}_{\mu,\nu}(\varphi \sum_{i=1}^{k} ||x_i||^s, r) \\ P^{'}_{\mu,\nu}(\varphi(\sum_{i=1}^{k} ||x_i||^{ks} + \Pi_{i=1}^{k}||x_i||^s), r) \end{cases}$$

for all $x_1, x_2, \cdots, x_k \in X$ and all $r > 0$, where φ, s are constants with $\varphi > 0$. Then there exists a unique cubic mapping $C : X \to Y$ such that

$$P_{\mu,\nu}(f(x) - C(x), r) \geq \begin{cases} P^{'}_{\mu,\nu}(\varphi, 4r \mid n^3 - 1 \mid r) & \\ P^{'}_{\mu,\nu}\left(\varphi||x||^s, 4r \mid n^3 - n^s \mid r\right) & ; s \neq 3 \\ P^{'}_{\mu,\nu}\left(\varphi||x||^{ks}, 4r \mid n^3 - n^{ks} \mid r\right) & ; s \neq \frac{3}{k} \end{cases}$$

for all $x \in X$ and all $r > 0$.

5 Stability Results for the Functional Eq. (1): Fixed Point Method

In this section, we establish the generalized Ulam-Hyers Stability of the functional Eq. (1) in Intuitionistic fuzzy Normed space via fixed point method.

For to prove the stability result, we define the following: η_i is a constant such that

$$\eta_i = \begin{cases} n & if \quad i = 0 \\ \frac{1}{n} & if \quad i = 1 \end{cases}$$

and Ω is the set such that $\Omega = \{g/g : X \to Y, g(0) = 0\}$.

Theorem 5.1. *Let $f : X \to Y$ be a mapping for which there exist a function $\alpha : X^3 \to Z$ with the condition*

$$\lim_{k\to\infty} P^{'}_{\mu,\nu}\left(\alpha(\eta_i^k x_1, \eta_i^k x_2, \cdots, \eta_i^k x_k), \eta_i^{3k} r\right) = 1_{L^*} \tag{47}$$

for all $x_1, x_2, \cdots, x_k \in X$ and $r > 0$ and satisfying the functional inequality

$$P_{\mu,\nu}\left(Df(x_1,x_2,\cdots,x_k), r\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\alpha(x_1,x_2,\cdots,x_k), r\right) \tag{48}$$

for all $x_1, x_2, \cdots, x_k \in X$ and $r > 0$. If there exists $L = L(i)$ such that the function $x \to \beta(x) = \frac{1}{4}\alpha\left(\frac{x}{n}, 0, \cdots, 0\right)$ has the property

$$P^{'}_{\mu,\nu}\left(L\frac{1}{\eta_i^3}\beta(\eta_i x), r\right) = P^{'}_{\mu,\nu}\left(\beta(x), r\right) \tag{49}$$

for all $X \in X$ and $r > 0$. Then there exists unique cubic function $C : X \to Y$ satisfying the functional Eq. (1) and

$$P_{\mu,\nu}\left(f(x) - C(x), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\beta(x), \frac{L^{1-i}}{1-L} r\right) \tag{50}$$

for all $x \in X$ and $r > 0$.

Proof. Let d be a general metric on Ω, such that

$$d(g,h) = \inf\{k \in (0,\infty) | P_{\mu,\nu}(g(x) - h(x), r) \geq_{L^*} P'_{\mu,\nu}\left(\beta(x), kr\right), x \in X, r > 0\}.$$

It is easy to see that (Ω, d) is complete. Define $T : \Omega \to \Omega$ by $Tg(x) = \frac{1}{\eta_i^3} g(\eta_i x)$, for all $x \in X$. For $g, h \in \Omega$, we have

$$d(g,h) \leq k$$

$$\Rightarrow P_{\mu,\nu}(g(x) - h(x), r) \geq_{L^*} P'_{\mu,\nu}(\beta(x), kr)$$

$$\Rightarrow P_{\mu,\nu}\left(\frac{g(\eta_i x)}{\eta_i^3} - \frac{h(\eta_i x)}{\eta_i^3}, r\right) \geq_{L^*} P'_{\mu,\nu}\left(\beta(\eta_i x), k\eta_i^3 r\right)$$

$$\Rightarrow P_{\mu,\nu}\left(Tg(x) - Th(x), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\beta(x), kLr\right)$$

$$\Rightarrow d\left(Tg(x), Th(x)\right) \leq kL$$

$$\Rightarrow d(Tg, Th) \leq Ld(g,h)$$

for all $g, h \in \Omega$. Therefore T is strictly contractive mapping on Ω with Lipschitz constant L. Replacing $(x_1, x_2, \cdots, x_k)$ by $(x, 0, \cdots, 0)$ in (48), we get

$$P_{\mu,\nu}(4f(nx) - 4n^3 f(x), r) \geq_{L^*} P'_{\mu,\nu}(\alpha(x, 0, \cdots, 0), r) \tag{51}$$

for all $x \in X$ and $r > 0$. using (IFN2) in (51), we arrive

$$P_{\mu,\nu}\left(\frac{f(nx)}{n^3} - f(x), r\right) \geq_{L^*} P'_{\mu,\nu}\left(\alpha(x, 0, \cdots, 0), 4n^3 r\right) \tag{52}$$

for all $x \in X$ and $r > 0$ with the help of (49) when $i = 0$, it follows from (52), we get

$$\Rightarrow P_{\mu,\nu}\left(\frac{f(nx)}{n^3}-f(x),r\right)\geq_{L^*} P^{'}_{\mu,\nu}\left(\beta(x),Lr\right) \tag{53}$$
$$\Rightarrow d(Tf,f)\leq L=L^1=L^{1-i}.$$

Replacing x by $\frac{x}{n}$ in (51), we attain

$$P_{\mu,\nu}\left(f(x)-n^3f\left(\frac{x}{n}\right),r\right)\geq_{L^*} P^{'}_{\mu,\nu}\left(\alpha\left(\frac{x}{n},0,\cdots,0\right),4r\right) \tag{54}$$

for all $x\in X$ and $r>0$ with the help of (49) when $i=1$, it follows from (54) we get

$$\Rightarrow P_{\mu,\nu}\left(f(x)-n^3f\left(\frac{x}{n}\right),r\right)\geq_{L^*} P^{'}_{\mu,\nu}\left(\beta(x),r\right) \tag{55}$$
$$\Rightarrow d(f,Tf)\leq 1=L^0=L^{1-i}.$$

Then from (53) and (55), we can conclude

$$d(f,Tf)\leq L^{1-i}<\infty.$$

Now from the fixed point alternative in both cases, it follows that there exists a fixed point C of T in Ω such that

$$\lim_{n\to\infty} P_{\mu,\nu}\left(\frac{f(\eta_i^n x)}{\eta_i^n}-C(x),r\right)\to 1_{L^*} \tag{56}$$

for all $x\in X$ and $r>0$. Replacing $(x_1,x_2,\cdots,x_k)$ by $(\eta_ix_1,\eta_ix_2,\cdots,\eta_ix_k)$ in (48), we obtain

$$P_{\mu,\nu}\left(\frac{1}{\eta_i^{3n}}Df(\eta_ix_1,\eta_ix_2,\cdots,\eta_ix_k),r\right)\geq_{L^*} P^{'}_{\mu,\nu}(\alpha(\eta_ix_1,\eta_ix_2,\cdots,\eta_ix_k),\eta_i^{3n}r) \tag{57}$$

for all $r>0$ and all $x_1,x_2,\cdots,x_k\in X$. By proceeding the same procedure as in the Theorem 4.1, we can prove the function $C:X\to Y$ satisfies the functional Eq. (1). By fixed point alternative, since C is unique fixed point of T in the set $\Delta=\{f\in\Omega|d(f,C)<\infty\}$. Therefore C is a unique function such that

$$P_{\mu,\nu}(f(x)-C(x),r)\geq_{L^*} P^{'}_{\mu,\nu}(\beta(x),kr) \tag{58}$$

for all $x\in X$ and $k,r>0$. Again using the fixed point alternative, we obtain

$$d(f,C)\leq\frac{1}{1-L}d(f,Tf)$$

$$\Rightarrow f(f,C) \leq \frac{L^{1-i}}{1-L}$$

$$\Rightarrow P_{\mu,\nu}(f(x)-C(x),r) \geq_{L^*} P'_{\mu,\nu}\left(\beta(x), \frac{L^{1-i}}{1-L}r\right),$$

for all $x \in X$ and $r > 0$. This completes the proof of the theorem.

From Theorem 5.1, we obtain the following corollary concerning the stability for the functional Eq. (1).

Corollary 5.1. *Suppose that a function $f : X \to Y$ satisfies the inequality*

$$P_{\mu,\nu}(Df(x_1,x_2,\cdots,x_k),r) \geq_{L^*} \begin{cases} P'_{\mu,\nu}(\varphi,r) \\ P'_{\mu,\nu}(\varphi\sum_{i=1}^{k}||x_i||^s,r) \\ P'_{\mu,\nu}(\varphi(\sum_{i=1}^{k}||x_i||^{ks}+\Pi_{i=1}^{3}||x_i||^s),r) \end{cases}$$

for all $x_1,x_2,\cdots,x_k \in X$ and $r > 0$, where φ, s are constants with $\varphi > 0$. Then there exists a unique cubic mapping $C : X \to Y$ such that

$$P_{\mu,\nu}(f(x)-C(x),r) \geq_{L^*} \begin{cases} P'_{\mu,\nu}(\varphi, 4r \mid n^3-1 \mid r) & \\ P'_{\mu,\nu}\left(\varphi||x||^s, 4r \mid n^3-n^s \mid r\right) & ; s \neq 3 \\ P'_{\mu,\nu}\left(\varphi||x||^{ks}, 4r \mid n^3-n^{ks} \mid r\right) & ; s \neq \frac{3}{k} \end{cases}$$

for all $x \in X$ and $r > 0$.

Proof. Setting

$$\alpha(x_1,x_2,\cdots,x_k) = \begin{cases} \varphi \\ \varphi(\sum_{i=1}^{k}||x_i||^s) \\ \varphi(\prod_{i=1}^{k}||x_i||^s+\sum_{i=1}^{k}||x_i||^{ks}) \end{cases}$$

for all $x_1,x_2,\cdots,x_k \in X$. Then

$$P'_{\mu,\nu}\left(\alpha\left(\eta_i^k x_1, \eta_i^k x_2, \cdots, \eta_i^k x_k\right), \eta_i^{3k} r\right) = \begin{cases} P'_{\mu,\nu}(\varphi, \eta_i^{3k} r) \\ P'_{\mu,\nu}\left(\varphi\sum_{i=1}^{k}||x_i||^s, \eta_i^{(3-s)k} r\right) \\ P'_{\mu,\nu}\left(\varphi(\sum_{i=1}^{k}||x_i||^{ks}+\Pi_{i=1}^{3}||x_i||^s), \eta_i^{(3-ks)k} r\right) \end{cases}$$

$$= \begin{cases} \longrightarrow \quad 1 \quad as \quad k \longrightarrow \infty, \\ \longrightarrow \quad 1 \quad as \quad k \longrightarrow \infty, \\ \longrightarrow \quad 1 \quad as \quad k \longrightarrow \infty. \end{cases}$$

Thus, (47) is holds. But we have

$$\beta(x)=\frac{1}{4}\alpha\left(\frac{x}{n},0,\cdots,0\right)$$

has the property

$$P'_{\mu,\nu}\left(L\frac{1}{\eta_i^3}\beta(\eta_i x),r\right)\geq P'_{\mu,\nu}(\beta(x),r)$$

for all $x\in X$ and $r>0$. Hence

$$P'_{\mu,\nu}(\beta(x),r)=N'\left(\alpha\left(\frac{x}{n},0,\cdots,0\right),4r\right)$$

$$=\begin{cases}P'_{\mu,\nu}(\varphi,4r)\\ P'_{\mu,\nu}\left(\varphi||\frac{x}{n}||^s,4r\right)\\ P'_{\mu,\nu}\left(\varphi||\frac{x}{n}||^{ks},4r\right).\end{cases}$$

Now,

$$P'_{\mu,\nu}\left(\frac{1}{\eta_i^3}\beta(\eta_i x),r\right)=\begin{cases}P'_{\mu,\nu}\left(\frac{\varphi}{\eta_i^3},4r\right)\\ P'_{\mu,\nu}\left(\frac{\varphi}{\eta_i^3}\left(\frac{1}{n^s}\right)||\eta_i x||^s,4r\right)\\ P'_{\mu,\nu}\left(\frac{\varphi}{\eta_i^3}\left(\frac{1}{n^{ns}}\right)||\eta_i x||^{ns},4r\right)\end{cases}$$

$$=\begin{cases}P'_{\mu,\nu}(\eta_i^{-3}\beta(x),4r)\\ P'_{\mu,\nu}(\eta_i^{s-3}\beta(x),4r)\\ P'_{\mu,\nu}(\eta_i^{ks-3}\beta(x),4r)\end{cases}$$

Now from the following cases for the conditions of η_i.

Case (i): $L=n^{-3}\quad for\quad s=0\quad if\quad i=0$

$$P_{\mu,\nu}(f(x)-C(x),r)\geq_{L^*}P'_{\mu,\nu}\left(\frac{L^{1-i}}{1-L}\beta(x),r\right)\geq_{L^*}P'_{\mu,\nu}\left(\frac{\varphi(n^{-3})}{1-n^{-3}},4r\right)\geq_{L^*}P'_{\mu,\nu}\left(\varphi,4(n^3-1)r\right)$$

Case (ii): $L=n^{3}\quad for\quad s=0\quad if\quad i=1$

$$P_{\mu,\nu}(f(x)-C(x),r)\geq_{L^*}P'_{\mu,\nu}\left(\frac{L^{1-i}}{1-L}\beta(x),r\right)\geq_{L^*}P'_{\mu,\nu}\left(\frac{\varphi}{1-n^3},4r\right)\geq_{L^*}P'_{\mu,\nu}\left(\varphi,4(1-n^3)r\right)$$

Case (iii): $L=n^{s-3}\quad for\quad s<3\quad if\quad i=0$

$$N(f(x)-C(x),r)\geq_{L^*}P'_{\mu,\nu}\left(\frac{L^{1-i}}{1-L}\beta(x),r\right)\geq_{L^*}P'_{\mu,\nu}\left(\frac{n^{s-3}}{1-n^{s-3}}\frac{\varphi||x||^s}{n^s},4r\right)$$
$$\geq_{L^*}P'_{\mu,\nu}\left(\varphi||x||^s,4r(n^3-n^s)\right)$$

Case (iv): $L = n^{3-s} \quad for \quad s > 3 \quad if \quad i = 1$

$$
\begin{aligned}
P_{\mu,\nu}(f(x) - C(x), r) &\geq_{L^*} P^{'}_{\mu,\nu}\left(\frac{L^{1-i}}{1-L}\beta(x), r\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\frac{1}{1-n^{3-s}}\frac{\varphi||x||^s}{n^s}, 4r\right) \\
&\geq_{L^*} P^{'}_{\mu,\nu}\left(\varphi||x||^s, 4r(n^s - n^3)\right)
\end{aligned}
$$

Case (v): $L = n^{ks-3} \quad for \quad s < \frac{3}{k} \quad if \quad i = 0$

$$
\begin{aligned}
P_{\mu,\nu}(f(x) - C(x), r) &\geq_{L^*} N'\left(\frac{L^{1-i}}{1-L}\beta(x), r\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\frac{n^{ks-3}}{1-n^{ks-3}}\frac{\varphi||x||^{ks}}{n^{ks}}, 4r\right) \\
&\geq_{L^*} P^{'}_{\mu,\nu}\left(\varphi||x||^{ks}, 4r(n^3 - n^{ks})\right)
\end{aligned}
$$

Case (vi): $L = n^{3-ks} \quad for \quad s < \frac{3}{k} \quad if \quad i = 1$

$$
\begin{aligned}
P_{\mu,\nu}(f(x) - C(x), r) &\geq_{L^*} P^{'}_{\mu,\nu}\left(\frac{L^{1-i}}{1-L}\beta(x), r\right) \geq_{L^*} P^{'}_{\mu,\nu}\left(\frac{1}{1-n^{3-ks}}\frac{\varphi||x||^{ks}}{n^{ks}}, 4r\right) \\
&\geq_{L^*} P^{'}_{\mu,\nu}\left(\varphi||x||^{ks}, 4r(n^{ks} - n^3)\right).
\end{aligned}
$$

Hence the proof is completed.

Acknowledgements This work was partially supported by FCT and CEMAT by the projects UIDB/04621/2020 and UIDP/04621/2020.

References

1. Bag, T., Samanta, S.K.: Fuzzy bounded linear operators. Fuzzy Sets Syst. **151**(3), 513–547 (2005)
2. Bag, T., Samanta, S.K.: Some fixed point theorems in fuzzy normed linear spaces. Inform. Sci. **177**(16), 3271–3289 (2007)
3. Bodaghi, A.: Cubic derivations on Banach algebras. Acta Math. Vietnam. **38**(4), 517–528 (2013)
4. Bodaghi, A.: Stability of a quartic functional equation. Sci. World J. **2014**, 9 (2014). Article ID 752146
5. Bodaghi, A.: Intuitionistic fuzzy stability of the generalized forms of cubic and quartic functional equations. J. Intel. Fuzzy Syst. **30**, 2309–2317 (2016)
6. Bodaghi, A., Alias, I.A., Ghahramani, M.H.: Ulam stability of a quartic functional equation. Abstr. Appl. Anal. **2012**, 9 (2012). Article ID 232630
7. Bodaghi, A., Moosavi, S.M., Rahimi, H.: The generalized cubic functional equation and the stability of cubic Jordan *-derivations. Ann. Univ. Ferrara Sez. VII Sci. Mat. **59**(2), 235–250 (2013)
8. Bodaghi, A., Zabandan, Gh: On the stability of quadratic (*-) derivations on (*-) Banach algebras. Thai J. Math. **12**(2), 343–356 (2014)
9. Cădariu, L., Radu, V.: Fixed points and the stability of quadratic functional equations. An. Univ. Timişoara Ser. Mat.-Inform. **41**(1), 25–48 (2003)

10. Cadariu, L., Radu, V.: On the stability of the Cauchy functional equation: a fixed point approach. Grazer Math. Ber. **346**, 43–52 (2004)
11. Cadariu, L., Radu, V.: Fixed points and stability for functional equations in probabilistic metric and random normed spaces. Fixed Point Theory Appl. **2009**, 18 (2009). Article ID 589143
12. Diaz, J.B., Margolis, B.: A fixed point theorem of the alternative for contractions on a generalized complete metric space. Bull. Am. Math. Soc. **74**, 305–309 (1968)
13. Fang, J.X.: On I-topology generated by fuzzy norm. Fuzzy Sets Syst. **157**(20), 2739–2750 (2006)
14. Hyers, D.H.: On the stability of the linear functional equation. Proc. Nat. Acad. Sci. USA **27**, 222–224 (1941)
15. Mohiuddine, S.A., Sevli, H.: Stability of pexiderized quadratic functional equation in intuitionistic fuzzy normed space. J. Comput. Appl. Math. **235**, 2137–214 (2011)
16. Mursaleen, M., Lohani, Q.M.D.: Intuitionistic fuzzy 2-normed space and some related concepts. Chaos Solitons Fractals **42**(1), 224–234 (2009)
17. Park, J.H.: Intuitionistic fuzzy metric spaces. Chaos Solitons Fractals **22**(5), 1039–1046 (2004)
18. Rassias, J.M., Eslamian, M.: Fixed points and stability of nonic functional equation in quasi-β-normed spaces. Contemp. Anal. Appl. Math. **3**(2), 293–309 (2015)
19. Saadati, R.: A note on "some results on the IF-normed spaces". Chaos Solitons Fractals **41**(1), 206–213 (2009)
20. Saadati, R., Cho, Y.J., Vahidi, J.: The stability of the quartic functional equation in various spaces. Comput. Math. Appl. **60**(7), 1994–2002 (2010)
21. Saadati, R., Park, C.: Non-archimedean L-fuzzy normed spaces and stability of functional equations. Comput. Math. Appl. **60**(8), 2488–2496 (2010)
22. Saadati, R., Park, J.H.: On the intuitionistic fuzzy topological spaces. Chaos Solitons Fractals **27**(2), 331–344 (2006)
23. Saadati, R., Razani, A., Adibi, H.: A common fixed point theorem in L-fuzzy metric spaces. Chaos Solitons Fractals **33**(2), 358–363 (2007)
24. Saadati, R., Sedghi, S., Shobe, N.: Modified intuitionistic fuzzy metric spaces and some fixed point theorems. Chaos Solitons Fractals **38**(1), 36–47 (2008)
25. Pinelas, S., Govindan, V., Tamilvanan, K.: Pfister 16-square quadratic functional equation. Global J. Math. **12**(1), 760–772 (2018)
26. Pinelas, S., Govindan, V., Tamilvanan, K.: Stability of cubic functional equation in random normed space. J. Adv. Math. **14**(2), 7864–7877 (2018)
27. Pinelas, S., Govindan, V., Tamilvanan, K.: Stability of a quartic functional equation. J. Fixed Point Theory Appl. **20**(4), 10 (2018)
28. Pinelas, S., Govindan, V., Tamilvanan, K.: Stability of non-additive functional equation. IOSR J. Math. **14**(2–I), 60–78 (2018)
29. Ulam, S.M.: Problems in Modern Mathematics, Chapter VI, Science edn. Wiley, New York (1940)
30. Xu, T.Z., Rassias, M.J., Xu, W.X.: Stability of a general mixed additive-cubic functional equation in non-archimedean fuzzy normed spaces. J. Math. Phys. **51**(9), 19 (2010). 093508
31. Xu, T.Z., Rassias, M.J., Xu, W.X., Rassias, J.M.: A fixed point approach to the intuitionistic fuzzy stability of quintic and sextic functional equations. Iran. J. Fuzzy Syst. **9**(5), 21–40 (2012)
32. Zadeh, L.A.: Fuzzy sets. Inform. Control **8**, 338–353 (1965)

An Abstract Impulsive Second-Order Functional-Differential Cauchy Problem with Nonlocal Conditions

Haydar Akça, Jamal Benbourenane, Valéry Covachev, and Zlatinka Covacheva

Abstract The main concern of the paper is to prove the existence, uniqueness and continuous dependence of mild and classical solutions of a semilinear impulsive second-order functional-differential equation with nonlocal initial conditions. We consider the Cauchy problem in general Banach spaces, and apply the theory of strongly continuous cosine families of linear operators and the Banach fixed-point theorem.

1 Introduction

Many evolutionary processes in nature are characterized by the fact that, at certain instants of time, they experience a rapid change of their states. The theory of the impulsive differential equations is one of the attractive branches of differential equations, which has extensive realistic mathematical modelling applications in physics, chemistry, engineering, and biological and medical sciences. The nonlocal condition generalizes the classical initial condition. In our previous papers [1, 2], we found sufficient conditions for the existence, uniqueness and continuous dependence of a mild solution of a first-order impulsive functional-differential evolution nonlocal Cauchy problem such that the linear part of the right-hand side of the differential equation is given by the infinitesimal generator of a strongly continuous semigroup of bounded linear operators.

H. Akça · J. Benbourenane
Abu Dhabi University, Abu Dhabi, UAE
e-mail: haydar.akca@adu.ac.ae

J. Benbourenane
e-mail: jamal.benbourenane@adu.ac.ae

V. Covachev (✉) · Z. Covacheva
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
e-mail: vcovachev@hotmail.com

Z. Covacheva
e-mail: zkovacheva@hotmail.com

Z. Covacheva
University of Mining and Geology, Sofia, Bulgaria

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_53

2 Statement of the Problem

We consider an impulsive abstract nonlocal semilinear second-order functional-differential Cauchy problem in the form

$$u''(t) + Au(t) = f(t, u(t), u(b_1(t)), \dots, u(b_m(t))), \; t \in (0, T] \setminus \bigcup_{k=1}^{\kappa} \{\tau_k\}, \quad (1)$$

$$\Delta u(\tau_k) = I_k(u(\tau_k)), \quad \Delta u'(\tau_k) = \overline{I}_k(u(\tau_k), u'(\tau_k)), \quad k = \overline{1, \kappa}, \quad (2)$$

$$u(0) = u_0, \quad u'(0) + \sum_{i=1}^{p} g_i u(t_i) = u_1, \quad (3)$$

where A is a linear operator from a real Banach space X into itself, $u : [0, T] \to X$, $f : [0, T] \times X^{m+1} \to X$, $b_i : [0, T] \to [0, T]$ $(i = \overline{1, m})$, $u_0, u_1 \in X$, $g_i \in \mathbb{R}$ $(i = \overline{1, p})$ and $\Delta u(\tau_k) = u(\tau_k + 0) - u(\tau_k - 0) \equiv u(\tau_k + 0) - u(\tau_k)$, $\Delta u'(\tau_k) = u'(\tau_k + 0) - u'(\tau_k - 0) \equiv u'(\tau_k + 0) - u'(\tau_k), 0 < \tau_1 < \tau_2 < \cdots < \tau_\kappa < T$ are the instants of impulse effect and $0 < t_1 < t_2 < \cdots < t_p < T$.

The main concern of the present paper is to find sufficient conditions for the existence, uniqueness and continuous dependence of mild and classical solutions of problem (1)–(3).

3 Preliminaries

We shall need the following definitions [6–9].

Definition 1. A one-parameter family $\{C(t) : \; t \in \mathbb{R}\}$ of bounded linear operators mapping the Banach space X into itself is called a *strongly continuous cosine family* if and only if

1. $C(s + t) + C(s - t) = 2C(s)C(t)$ for all $s, t \in \mathbb{R}$.
2. $C(0) = I$ (the identity operator).
3. $C(t)x$ is continuous in t on $\mathbb{R}$ for each fixed $x \in X$.

Definition 2. The infinitesimal generator of a strongly continuous cosine family $\{C(t)\}$ is the operator $A : \; X \supset \mathscr{D}(A) \to X$ defined by

$$Ax := \left. \frac{d^2}{dt^2} C(t)x \right|_{t=0}, \quad x \in \mathscr{D}(A),$$

where

$$\mathscr{D}(A) := \left\{ x \in X : \; C(t)x \; \text{ is of class } \; C^2 \; \text{ with respect to } \; t \right\}.$$

We introduce the assumptions:

A1. The operator $-A$ is the infinitesimal generator of a strongly continuous cosine family $\{C(t) : t \in \mathbb{R}\}$ of bounded linear operators from X to itself.
A2. The adjoint operator A^* is densely defined in X^*, *i.e.*, $\overline{\mathscr{D}(A^*)} = X^*$.

Let us denote

$$E := \left\{x \in X : \ C(t)x \ \text{ is of class } \ C^1 \ \text{ with respect to } \ t\right\}.$$

The associated sine family $\{S(t) : t \in \mathbb{R}\}$ is defined by

$$S(t)x := \int_0^t C(s)x\, ds, \qquad x \in X, \quad t \in \mathbb{R}.$$

Further on, we denote by $\|C(t)\|$, $\|S(t)\|$ and $\|A\|$ the operator norms of $C(t)$, $S(t)$ and A in the Banach space X, respectively. From Assumption **A1**, it follows that there exists a constant $M \geq 1$ such that

$$\|C(t)\| \leq M \quad \text{and} \quad \|S(t)\| \leq M \quad \text{for} \quad t \in [0, T]. \tag{4}$$

Following [5], we present a result obtained by J. Bochenek in [4].

Let us consider the Cauchy problem

$$\begin{aligned} &x''(t) + Ax(t) = h(t), \quad t \in (0, T], \\ &x(0) = x_0, \qquad x'(0) = x_1. \end{aligned} \tag{5}$$

Definition 3. A function $x : [0, T] \to X$ is said to be a *classical solution* of problem (5) if

$$\begin{aligned} &x \in C^1([0, T], X) \cap C^2((0, T], X), \\ &x(0) = x_0 \quad \text{and} \quad x'(0) = x_1, \\ &x''(t) + Ax(t) = h(t) \quad \text{for} \quad t \in (0, T]. \end{aligned}$$

Theorem 1. *Suppose that*

1. *Assumptions* **A1** *and* **A2** *are satisfied;*
2. $h : [0, T] \to X$ *is Lipschitz continuous;*
3. $x_0 \in \mathscr{D}(A)$ *and* $x_1 \in E$.

Then, problem (5) *has a unique classical solution given by the formula*

$$x(t) = C(t)x_0 + S(t)x_1 + \int_0^t S(t-s)h(s)\,ds, \quad t \in [0,T].$$

It is easy to see that this result can be generalized for the impulsive system

$$x''(t) + Ax(t) = h(t), \quad t \in (0,T] \setminus \{\tau_1, \tau_2, \ldots, \tau_\kappa\}, \tag{6}$$

$$\Delta x(\tau_k) = I_k, \quad \Delta x'(\tau_k) = \overline{I}_k, \quad k = \overline{1,\kappa}, \tag{7}$$

$$x(0) = x_0, \quad x'(0) = x_1. \tag{8}$$

For convenience, we denote $J = [0,T]$, $J_0 = [0,\tau_1]$, $J_k = (\tau_k, \tau_{k+1}]$, $k = \overline{1, \kappa - 1}$, $J_\kappa = (\tau_\kappa, T]$, $J' = J \setminus \{0, \tau_1, \tau_2, \ldots, \tau_\kappa\}$. For a function $x : J \to X$, we denote by $x_{|k}$ the restriction of x to J_k, $k = \overline{0,\kappa}$, with $\|x_{|k}\|_{J_k} = \sup\limits_{s \in J_k} \|x_{|k}(s)\|$. If the function $x_{|k}$ happens to be differentiable, we denote its derivative by $x'_{|k}$. Further we introduce the following classes of functions:

$$\begin{aligned} PC(J', X) = \big\{ x : J \to X \mid x_{|k} \in C(J_k, X),\ k = \overline{0,\kappa}, \\ \text{and there exist } x(\tau_k + 0),\ k = \overline{1,\kappa} \big\}, \\ PC^1(J', X) = \big\{ x \in PC(J', X) \mid x'_{|k} \in C(J_k, X),\ k = \overline{0,\kappa}, \\ \text{and there exist } x'(\tau_k + 0),\ k = \overline{1,\kappa} \big\}. \end{aligned}$$

$PC(J', X)$ is a Banach space with norm $\|x\|_{PC} = \max\big\{\|x_{|k}\|_{J_k},\ k = \overline{0,\kappa}\big\}$, and $PC^1(J', X)$ is a Banach space with norm $\|x\|_{PC^1} = \|x\|_{PC} + \|x'\|_{PC}$.

Definition 4. A function $x \in PC^1(J', X) \cap C^2(J', X)$ is called a *classical solution* of problem (6)–(8) if it satisfies the differential equation (6) on J', together with the impulse conditions (7) and the initial conditions (8).

Theorem 2. [3] *Suppose that*

1. *Assumptions* **A1** *and* **A2** *are satisfied;*
2. $h \in PC(J', X)$ *is such that its restrictions to* J_k *are Lipschitz continuous,* $k = \overline{0,\kappa}$*;*
3. $x_0 \in \mathscr{D}(A)$ *and* $x_1 \in E$.
4. $I_k \in \mathscr{D}(A)$ *and* $\overline{I}_k \in E$ *for* $k = \overline{1,\kappa}$.

Then, problem (6)–(8) *has a unique classical solution given by the formula*

$$\begin{aligned} x(t) = C(t)x_0 + S(t)x_1 + \int_0^t S(t-s)h(s)\,ds \\ + \sum_{0<\tau_k<t} C(t-\tau_k)I_k + \sum_{0<\tau_k<t} S(t-\tau_k)\overline{I}_k, \quad t \in J. \end{aligned} \tag{9}$$

Theorem 2 can be proved by applying Theorem 1 on each interval of continuity J_k, $k = \overline{0, \kappa}$.

This theorem suggests the following definition.

Definition 5. A function $u \in PC^1(J', X)$ satisfying the integro-summary equation

$$u(t) = C(t)u_0 + S(t)\left(u_1 - \sum_{i=1}^{p} g_i u(t_i)\right) \tag{10}$$

$$+ \int_0^t S(t-s) f(s, u(s), u(b_1(s)), \dots, u(b_m(s)))\, ds$$

$$+ \sum_{0<\tau_k<t} C(t-\tau_k) I_k(u(\tau_k)) + \sum_{0<\tau_k<t} S(t-\tau_k) \overline{I}_k(u(\tau_k), u'(\tau_k)), \quad t \in J,$$

is said to be a *mild solution* of the nonlocal problem (1)–(3).

4 Main Results

4.1 Existence and Uniqueness of a Mild Solution

Theorem 3. *Suppose that*

1. *Assumption* **A1** *is satisfied;*
2. *The function $t \mapsto f(t, x, y_1, \dots, y_m)$ belongs to $PC(J', X)$, and there exists a positive constant L_1 such that*

$$\|f(t, x, y_1, \dots, y_m) - f(t, \tilde{x}, \tilde{y}_1, \dots, \tilde{y}_m)\| \le L_1 \left(\|x - \tilde{x}\| + \sum_{j=1}^{m} \|y_j - \tilde{y}_j\|\right)$$

   for $t \in [0, T]$, $x, \tilde{x}, y_j, \tilde{y}_j \in X$, $j = \overline{1, m}$;
3. *$I_k : X \to E$ and $\overline{I}_k : X^2 \to X$ and there exist positive constants L_2 and L_3 such that*

$$\|I_k(x) - I_k(\tilde{x})\| \le L_2 \|x - \tilde{x}\| \quad \textit{and} \quad \|\overline{I}_k(x, y) - \overline{I}_k(\tilde{x}, \tilde{y})\| \le L_3(\|x - \tilde{x}\| + \|y - \tilde{y}\|)$$

   for $k = \overline{1, \kappa}$, $x, \tilde{x}, y, \tilde{y} \in X$;

4. $q := 2\tilde{M}\left[\sum_{i=1}^{p}|g_i| + (m+1)TL_1 + \kappa(L_2+L_3)\right] < 1,$ *where* $\tilde{M} = \max\{M, M'\}$, *$M$ was defined in* (4), *and* $M' = \sup\{\|C'(t)\| : t \in [0,T]\}$;
5. $u_0 \in E$ *and* $u_1 \in X$.

Then, problem (1)–(3) *has a unique mild solution.*

Proof. We can write Eq. (10) in an operator form

$$u = \mathscr{F}u,$$

where the operator $\mathscr{F} : PC^1(J', X) \to PC^1(J', X)$ is defined by

$$\begin{aligned}(\mathscr{F}u)(t) &= C(t)u_0 + S(t)\left(u_1 - \sum_{i=1}^{p} g_i u(t_i)\right)\\ &+ \int_0^t S(t-s)f(s, u(s), u(b_1(s)), \ldots, u(b_m(s)))\,ds\\ &+ \sum_{0<\tau_k<t} C(t-\tau_k)I_k(u(\tau_k)) + \sum_{0<\tau_k<t} S(t-\tau_k)\overline{I}_k(u(\tau_k), u'(\tau_k)), \quad t \in [0,T].\end{aligned}$$

Now, we show that $\mathscr{F}$ is a contraction on the Banach space $PC^1(J', X)$. In fact, for $u, \tilde{u} \in PC^1(J', X)$, we have

$$\begin{aligned}(\mathscr{F}u)(t) - (\mathscr{F}\tilde{u})(t) &= -S(t)\sum_{i=1}^{p} g_i(u(t_i) - \tilde{u}(t_i))\\ &+ \int_0^t S(t-s)\,(f(s, u(s), u(b_1(s)), \ldots, u(b_m(s)))\\ &\qquad - f(s, \tilde{u}(s), \tilde{u}(b_1(s)), \ldots, \tilde{u}(b_m(s))))\,ds\\ &+ \sum_{0<\tau_k<t} C(t-\tau_k)\,(I_k(u(\tau_k)) - I_k(\tilde{u}(\tau_k)))\\ &+ \sum_{0<\tau_k<t} S(t-\tau_k)\left(\overline{I}_k(u(\tau_k), u'(\tau_k)) - \overline{I}_k(\tilde{u}(\tau_k), \tilde{u}'(\tau_k))\right), \quad t \in [0,T],\end{aligned}$$

hence,

$$
\begin{aligned}
&\|(\mathscr{F}u)(t) - (\mathscr{F}\tilde{u})(t)\| \le \|S(t)\| \sum_{i=1}^{p} |g_i| \cdot \|u(t_i) - \tilde{u}(t_i)\| \\
&+ \int_0^t \|S(t-s)\| \cdot \|f(s, u(s), u(b_1(s)), \ldots, u(b_m(s))) \\
&\qquad - f(s, \tilde{u}(s), \tilde{u}(b_1(s)), \ldots, \tilde{u}(b_m(s)))\| \, ds \\
&+ \sum_{0<\tau_k<t} \|C(t-\tau_k)\| \cdot \|I_k(u(\tau_k)) - I_k(\tilde{u}(\tau_k))\| \\
&+ \sum_{0<\tau_k<t} \|S(t-\tau_k)\| \cdot \|\overline{I}_k(u(\tau_k), u'(\tau_k)) - \overline{I}_k(\tilde{u}(\tau_k), \tilde{u}'(\tau_k))\| \\
&\le M \sum_{i=1}^{p} |g_i| \|u - \tilde{u}\|_{PC} + ML_1 \int_0^t \left(\|u(s) - \tilde{u}(s)\| + \sum_{j=1}^{m} \|u(b_j(s)) - \tilde{u}(b_j(s))\| \right) ds \\
&+ M \sum_{0<\tau_k<t} \left\{ L_2 \|u(\tau_k) - \tilde{u}(\tau_k)\| + L_3 \left(\|u(\tau_k) - \tilde{u}(\tau_k)\| + \|u'(\tau_k) - \tilde{u}'(\tau_k)\| \right) \right\} \\
&\le M \left\{ \left[\sum_{i=1}^{p} |g_i| + (m+1)TL_1 + \kappa(L_2 + L_3) \right] \|u - \tilde{u}\|_{PC} + \kappa L_3 \|u' - \tilde{u}'\|_{PC} \right\}. \qquad (11)
\end{aligned}
$$

Similarly, we obtain

$$
\begin{aligned}
(\mathscr{F}u)'(t) - (\mathscr{F}\tilde{u})'(t) &= -C(t) \sum_{i=1}^{p} g_i (u(t_i) - \tilde{u}(t_i)) \\
&+ \int_0^t C(t-s) \left(f(s, u(s), u(b_1(s)), \ldots, u(b_m(s))) \right. \\
&\qquad \left. - f(s, \tilde{u}(s), \tilde{u}(b_1(s)), \ldots, \tilde{u}(b_m(s))) \right) ds \\
&+ \sum_{0<\tau_k<t} C'(t-\tau_k) \left(I_k(u(\tau_k)) - I_k(\tilde{u}(\tau_k)) \right) \\
&+ \sum_{0<\tau_k<t} C(t-\tau_k) \left(\overline{I}_k(u(\tau_k), u'(\tau_k)) - \overline{I}_k(\tilde{u}(\tau_k), \tilde{u}'(\tau_k)) \right), \quad t \in [0, T],
\end{aligned}
$$

hence,

$$
\begin{aligned}
&\|(\mathscr{F}u)'(t) - (\mathscr{F}\tilde{u})'(t)\| \qquad (12) \\
&\le \left\{ M'\kappa L_2 + M \left[\sum_{i=1}^{p} |g_i| + (m+1)TL_1 + \kappa L_3 \right] \right\} \|u - \tilde{u}\|_{PC} + M\kappa L_3 \|u' - \tilde{u}'\|_{PC}.
\end{aligned}
$$

From (11) and (12), we derive the estimate

$$\begin{aligned}\|\mathscr{F}u-\mathscr{F}\tilde{u}\|_{PC^1} \le &\left\{M\left[2\sum_{i=1}^{p}|g_i|+2(m+1)TL_1+\kappa(L_2+2L_3)\right]\right.\\ &\left.+M'\kappa L_2\right\}\|u-\tilde{u}\|_{PC}+2M\kappa L_3\|u'-\tilde{u}'\|_{PC}\\ \le 2\tilde{M}&\left[\sum_{i=1}^{p}|g_i|+(m+1)TL_1+\kappa(L_2+L_3)\right]\|u-\tilde{u}\|_{PC^1}=q\|u-\tilde{u}\|_{PC^1}.\end{aligned}$$

Thus, the contraction mapping $\mathscr{F}$ has a unique fixed point $u\in PC^1(J',X)$, which is the mild solution of problem (1)–(3). □

Remark 1. In [3], for a system similar to (1)–(3), we proved the existence of a mild solution under considerably less restrictive conditions.

4.2 Existence and Uniqueness of a Classical Solution

Now let us consider system (1)–(3) satisfying assumptions **A1**, **A2**, and conditions 2, 3, 4 and 5 of Theorem 3 replaced respectively by

2′. The function $t\mapsto f(t,x,y_1,\dots,y_m)$ belongs to $PC(J',X)$ and there exists a positive constant $\tilde{L}_1$ such that

$$\begin{aligned}&\|f(t,x,y_1,\dots,y_m)-f(\tilde{t},\tilde{x},\tilde{y}_1,\dots,\tilde{y}_m)\|\\ &\le\tilde{L}_1\left(|t-\tilde{t}|+\|x-\tilde{x}\|+\sum_{j=1}^{m}\|y_j-\tilde{y}_j\|\right)\end{aligned}$$

for $t,\tilde{t}\in J_k$, $k=\overline{0,\kappa}$, $x,\tilde{x},y_j,\tilde{y}_j\in X$, $j=\overline{1,m}$;

3′. $I_k: X\to\mathscr{D}(A)$ and $\overline{I}_k: X^2\to E$, and there exist positive constants L_2 and L_3 such that

$$\|I_k(x)-I_k(\tilde{x})\|\le L_2\|x-\tilde{x}\| \quad\text{and}\quad \|\overline{I}_k(x,y)-\overline{I}_k(\tilde{x},\tilde{y})\|\le L_3(\|x-\tilde{x}\|+\|y-\tilde{y}\|)$$

for $k=\overline{1,\kappa}$, $x,\tilde{x},y,\tilde{y}\in X$;

4′. $\tilde{q}:=2\tilde{M}\left[\sum_{i=1}^{p}|g_i|+(m+1)T\tilde{L}_1+\kappa(L_2+L_3)\right]<1$;

5′. $u_0\in\mathscr{D}(A)$ and $u_1\in E$.

Next, we introduce the assumption

A3. The functions b_i $(i = \overline{1, m})$ are one-to-one $[0, T] \to [0, T]$, and satisfy $b_i(\tau_k) = \tau_k$ $(k = \overline{1, \kappa})$ and

$$|b_i(t) - b_i(\tilde{t})| \le \beta_i |t - \tilde{t}| \quad \text{for} \quad t, \tilde{t} \in J_k,$$

where β_i $(i = \overline{1, m})$ are some positive constants.

Similarly to Definition 4, we give the following definition.

Definition 6. A function $u \in PC^1(J', X) \cap C^2(J', X)$ is called a *classical solution* of problem (1)–(3) if it satisfies the differential equation (1) on J', together with the impulse conditions (2) and the nonlocal initial conditions (3).

Theorem 4. *Suppose that system* (1)–(3) *satisfies assumptions* **A1**–**A3***, as well as conditions* $2'$–$5'$*. Then, problem* (1)–(3) *has a unique classical solution.*

Proof. Since all assumptions of Theorem 3 are satisfied, problem (1)–(3) has a unique mild solution u. We shall show that u is a classical solution of problem (1)–(3).

Let $t, \tilde{t} \in J_k$ for some $k \in \{0, 1, \dots, \kappa\}$. Then, we have

$$u(t) - u(\tilde{t}) = \int_0^1 \frac{\partial}{\partial s} u(st + (1-s)\tilde{t})\, ds = \int_0^1 (t - \tilde{t}) u'(st + (1-s)\tilde{t})\, ds,$$

which implies

$$\|u(t) - u(\tilde{t})\| \le |t - \tilde{t}| \int_0^1 \|u'(st + (1-s)\tilde{t})\|\, ds \le |t - \tilde{t}| \sup_{s \in J} \|u'(s)\|.$$

Further on, by virtue of assumption **A3**, we obtain

$$\|u(b_i(t)) - u(b_i(\tilde{t}))\| \le |b_i(t) - b_i(\tilde{t})| \sup_{s \in J} \|u'(s)\| \le \beta_i |t - \tilde{t}| \sup_{s \in J} \|u'(s)\|$$

and, in view of condition $2'$,

$$\begin{aligned} &\|f(t, u(t), u(b_1(t)), \dots, u(b_m(t)) - f(\tilde{t}, u(\tilde{t}), u(b_1(\tilde{t}), \dots, u(b_m(\tilde{t})))\| \\ &\le \tilde{L}_1 \left\{ |t - \tilde{t}| + \|u(t) - u(\tilde{t})\| + \sum_{i=1}^m \|u(b_i(t)) - u(b_i(\tilde{t}))\| \right\} \\ &\le \tilde{L}_1 \left\{ 1 + \sup_{s \in J} \|u'(s)\| \left(1 + \sum_{i=1}^m \beta_i \right) \right\} |t - \tilde{t}|. \end{aligned}$$

This inequality shows that the mapping $[0, T] \ni t \mapsto f(t, u(t), u(b_1(t)), \dots, u(b_m(t))) \in X$ is Lipschitz continuous on each interval J_k, $k = \overline{0, \kappa}$. Thus, in view of Theorem 2, the problem

$$v''(t) + Av(t) = f(t, u(t), u(b_1(t)), \ldots, u(b_m(t))), \; t \in J',$$
$$\Delta v(\tau_k) = I_k(u(\tau_k)), \quad \Delta v'(\tau_k) = \overline{I}_k(u(\tau_k), u'(\tau_k)), \quad k = \overline{1, \kappa},$$
$$v(0) = u_0, \quad v'(0) + \sum_{i=1}^{p} g_i u(t_i) = u_1,$$

has a unique classical solution $v(t)$ given by the formula

$$v(t) = C(t)u_0 + S(t)\left(u_1 - \sum_{i=1}^{p} g_i u(t_i)\right)$$
$$+ \int_0^t S(t-s) f(s, u(s), u(b_1(s)), \ldots, u(b_m(s)))\, ds$$
$$+ \sum_{0<\tau_k<t} C(t-\tau_k) I_k(u(\tau_k)) + \sum_{0<\tau_k<t} S(t-\tau_k)\overline{I}_k(u(\tau_k), u'(\tau_k)), \quad t \in J.$$

Since the unique mild solution $u(t)$ of problem (1)–(3) satisfies Eq. (10), then $u(t) \equiv v(t)$, that is, $u(t)$ is the unique classical solution of problem (1)–(3). □

4.3 Continuous Dependence of a Solution on the Initial Condition and Bounded Perturbations of the Impulse Operators

First, we study the continuous dependence of a solution on the initial condition.

Theorem 5. *Let all assumptions of Theorem 3 be satisfied. Suppose that u and $\tilde{u}$ are mild solutions respectively of problem* (1)–(3) *and the impulsive system* (1), (2) *with initial condition*

$$\tilde{u}(0) = \tilde{u}_0, \quad \tilde{u}'(0) + \sum_{i=1}^{p} g_i \tilde{u}(t_i) = \tilde{u}_1, \qquad (\tilde{3})$$

where $\tilde{u}_0 \in \mathscr{D}(A)$ and $\tilde{u}_1 \in E$. Then,

$$\|u - \tilde{u}\|_{PC^1} \le \frac{2\tilde{M}}{1-q}\left(\|u_0 - \tilde{u}_0\| + \|u_1 - \tilde{u}_1\|\right). \qquad (13)$$

Proof. From Eq. (10) applied to the mild solutions u and $\tilde{u}$, we have

$$u(t) - \tilde{u}(t) = C(t)(u_0 - \tilde{u}_0) + S(t)\left[(u_1 - \tilde{u}_1) - \sum_{i=1}^{p} g_i(u(t_i) - \tilde{u}(t_i))\right]$$

$$
\begin{aligned}
&+ \int_0^t S(t-s)\,(f(s,u(s),u(b_1(s)),\dots,u(b_m(s))) \\
&\qquad\qquad - f(s,\tilde{u}(s),\tilde{u}(b_1(s)),\dots,\tilde{u}(b_m(s))))\,ds \\
&+ \sum_{0<\tau_k<t} C(t-\tau_k)(I_k(u(\tau_k)) - I_k(\tilde{u}(\tau_k))) \\
&+ \sum_{0<\tau_k<t} S(t-\tau_k)(\overline{I}_k(u(\tau_k),u'(\tau_k)) - \overline{I}_k(\tilde{u}(\tau_k),\tilde{u}'(\tau_k))), \quad t\in[0,T], \quad (14)
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\|u(t)-\tilde{u}(t)\| \le M\,(\|u_0-\tilde{u}_0\| + \|u_1-\tilde{u}_1\|) \qquad (15)\\
&+ M\left\{\left[\sum_{i=1}^{p}|g_i| + (m+1)TL_1 + \kappa(L_2+L_3)\right]\|u-\tilde{u}\|_{PC} + \kappa L_3\|u'-\tilde{u}'\|_{PC}\right\}.
\end{aligned}
$$

Similarly, we obtain

$$
\begin{aligned}
u'(t)-\tilde{u}'(t) &= C'(t)(u_0-\tilde{u}_0) + C(t)\left[(u_1-\tilde{u}_1) - \sum_{i=1}^{p} g_i(u(t_i)-\tilde{u}(t_i))\right] \\
&+ \int_0^t C(t-s)\,(f(s,u(s),u(b_1(s)),\dots,u(b_m(s))) \\
&\qquad\qquad - f(s,\tilde{u}(s),\tilde{u}(b_1(s)),\dots,\tilde{u}(b_m(s))))\,ds \\
&+ \sum_{0<\tau_k<t} C'(t-\tau_k)\,(I_k(u(\tau_k)) - I_k(\tilde{u}(\tau_k))) \\
&+ \sum_{0<\tau_k<t} C(t-\tau_k)\left(\overline{I}_k(u(\tau_k),u'(\tau_k)) - \overline{I}_k(\tilde{u}(\tau_k),\tilde{u}'(\tau_k))\right), \quad t\in[0,T],
\end{aligned}
$$

hence,

$$
\begin{aligned}
&\|u'(t)-\tilde{u}'(t)\| \le M'\|u_0-\tilde{u}_0\| + M\|u_1-\tilde{u}_1\| \qquad (16)\\
&+ \left\{M'\kappa L_2 + M\left[\sum_{i=1}^{p}|g_i| + (m+1)TL_1 + \kappa L_3\right]\right\}\|u-\tilde{u}\|_{PC} + M\kappa L_3\|u'-\tilde{u}'\|_{PC}.
\end{aligned}
$$

Adding together inequalities (15) and (16), we obtain

$$
\begin{aligned}
&\|u(t)-\tilde{u}(t)\| + \|u'(t)-\tilde{u}'(t)\| \le 2\tilde{M}\,(\|u_0-\tilde{u}_0\| + \|u_1-\tilde{u}_1\|) \\
&+ 2\tilde{M}\left[\sum_{i=1}^{p}|g_i| + (m+1)TL_1 + \kappa(L_2+L_3)\right]\|u-\tilde{u}\|_{PC^1} \\
&\equiv 2\tilde{M}\,(\|u_0-\tilde{u}_0\| + \|u_1-\tilde{u}_1\|) + q\|u-\tilde{u}\|_{PC^1}.
\end{aligned}
$$

Taking the supremum of the left-hand side over J, we deduce

$$\|u-\tilde{u}\|_{PC^1} \le 2\tilde{M}\left(\|u_0-\tilde{u}_0\|+\|u_1-\tilde{u}_1\|\right)+q\|u-\tilde{u}\|_{PC^1},$$

which implies estimate (13). □

This theorem shows that the unique mild solution u of problem (1)–(3) provided by Theorem 3 depends continuously on the initial data u_0, u_1. If u is the unique classical solution of problem (1)–(3) provided by Theorem 4, it is also a mild solution. So it satisfies estimate (13), consequently, it depends continuously on the initial data u_0, u_1.

Finally, we study the continuous dependence of a solution on the initial condition and bounded perturbations of the impulse operators.

Theorem 6. *Let all assumptions of Theorem 3 be satisfied. Suppose that u and $\tilde{u}$ are mild solutions respectively of problem* (1)–(3) *and system* (1) *provided with the impulse conditions*

$$\Delta\tilde{u}(\tau_k)=\tilde{I}_k(\tilde{u}(\tau_k)), \quad \Delta\tilde{u}'(\tau_k)=\tilde{\tilde{\bar{I}}}_k(\tilde{u}(\tau_k),\tilde{u}'(\tau_k)), \quad k=\overline{1,\kappa}, \tag{$\tilde{2}$}$$

and with initial condition ($\tilde{3}$), where $\tilde{u}_0 \in \mathscr{D}(A)$ and $\tilde{u}_1 \in E$. Here, $\tilde{I}_k : X \to E$ and $\tilde{\tilde{\bar{I}}}_k : X^2 \to X$ satisfy

$$\|\tilde{I}_k(x)-\tilde{I}_k(\tilde{x})\| \le L_2\|x-\tilde{x}\| \quad \text{and} \quad \|\tilde{\tilde{\bar{I}}}_k(x,y)-\tilde{\tilde{\bar{I}}}_k(\tilde{x},\tilde{y})\| \le L_3(\|x-\tilde{x}\|+\|y-\tilde{y}\|)$$

for $k=\overline{1,\kappa}$, $x,\tilde{x},y,\tilde{y} \in X$; moreover,

$$\|I_k-\tilde{I}_k\| := \sup_{x\in X}\|I_k(x)-\tilde{I}_k(x)\| < \infty \quad \text{and} \quad \|\bar{I}_k-\tilde{\tilde{\bar{I}}}_k\| := \sup_{(x,y)\in X^2}\|\bar{I}_k(x,y)-\tilde{\tilde{\bar{I}}}_k(x,y)\| < \infty.$$

Then,

$$\|u-\tilde{u}\|_{PC^1} \le \frac{2\tilde{M}}{1-q}\left[\|u_0-\tilde{u}_0\|+\|u_1-\tilde{u}_1\|+\sum_{k=1}^{\kappa}\left(\|I_k-\tilde{I}_k\|+\|\bar{I}_k-\tilde{\tilde{\bar{I}}}_k\|\right)\right]. \tag{17}$$

Proof. It suffices just to slightly modify the proof of Theorem 5. Instead of (14), we have

$$\begin{aligned}
u(t)-\tilde{u}(t) &= C(t)(u_0-\tilde{u}_0) + S(t)\left[(u_1-\tilde{u}_1) - \sum_{i=1}^{p} g_i(u(t_i)-\tilde{u}(t_i))\right] \\
&\quad + \int_0^t S(t-s)\,(f(s,u(s),u(b_1(s)),\ldots,u(b_m(s))) \\
&\qquad\qquad - f(s,\tilde{u}(s),\tilde{u}(b_1(s)),\ldots,\tilde{u}(b_m(s))))\,ds \\
&\quad + \sum_{0<\tau_k<t} C(t-\tau_k)(I_k(u(\tau_k)) - \tilde{I}_k(\tilde{u}(\tau_k))) \\
&\quad + \sum_{0<\tau_k<t} S(t-\tau_k)(\overline{I}_k(u(\tau_k),u'(\tau_k)) - \tilde{\overline{I}}_k(\tilde{u}(\tau_k),\tilde{u}'(\tau_k))), \quad t\in[0,T].
\end{aligned} \tag{18}$$

Next, we use the inequalities

$$\begin{aligned}
\|I_k(u(\tau_k)) - \tilde{I}_k(\tilde{u}(\tau_k))\| &\le \|I_k(u(\tau_k)) - I_k(\tilde{u}(\tau_k))\| + \|I_k(\tilde{u}(\tau_k)) - \tilde{I}_k(\tilde{u}(\tau_k))\| \\
&\le L_2\|u-\tilde{u}\|_{PC} + \|I_k - \tilde{I}_k\|
\end{aligned}$$

and, similarly,

$$\|\overline{I}_k(u(\tau_k),u'(\tau_k)) - \tilde{\overline{I}}_k(\tilde{u}(\tau_k),\tilde{u}'(\tau_k))\| \le L_3\left(\|u-\tilde{u}\|_{PC} + \|u'-\tilde{u}'\|_{PC}\right) + \|\overline{I}_k - \tilde{\overline{I}}_k\|.$$

In view of these inequalities, (18) implies

$$\begin{aligned}
\|u(t)-\tilde{u}(t)\| \le M\Bigg\{ &\|u_0-\tilde{u}_0\| + \|u_1-\tilde{u}_1\| + \sum_{k=1}^{\kappa}\left(\|I_k-\tilde{I}_k\| + \|\overline{I}_k - \tilde{\overline{I}}_k\|\right) \\
+ &\left[\sum_{i=1}^{p}|g_i| + (m+1)TL_1 + \kappa(L_2+L_3)\right]\|u-\tilde{u}\|_{PC} + \kappa L_3\|u'-\tilde{u}'\|_{PC}\Bigg\}.
\end{aligned} \tag{19}$$

Similarly, instead of (16), we obtain

$$\begin{aligned}
&\|u'(t)-\tilde{u}'(t)\| \\
&\le M'\left(\|u_0-\tilde{u}_0\| + \sum_{k=1}^{\kappa}\|I_k-\tilde{I}_k\|\right) + M\left(\|u_1-\tilde{u}_1\| + \sum_{k=1}^{\kappa}\|\overline{I}_k - \tilde{\overline{I}}_k\|\right) \\
&+ \left\{M'\kappa L_2 + M\left[\sum_{i=1}^{p}|g_i| + (m+1)TL_1 + \kappa L_3\right]\right\}\|u-\tilde{u}\|_{PC} + M\kappa L_3\|u'-\tilde{u}'\|_{PC}.
\end{aligned} \tag{20}$$

Adding together inequalities (19) and (20), and taking the supremum of the left-hand side over J, we deduce

$$\begin{aligned}\|u-\tilde{u}\|_{PC^1} &\le q\|u-\tilde{u}\|_{PC^1}\\ &+ 2\tilde{M}\left[\|u_0-\tilde{u}_0\|+\|u_1-\tilde{u}_1\|+\sum_{k=1}^{\kappa}\left(\|I_k-\tilde{I}_k\|+\|\overline{I}_k-\tilde{\overline{I}}_k\|\right)\right],\end{aligned}$$

which implies estimate (17). □

The last theorem was added following a referee's suggestion.

5 Conclusion

In the present paper, we considered a nonlocal Cauchy problem for a semilinear impulsive second-order functional-differential equation in a general Banach space. Under the assumption that the linear part of the equation is given by the infinitesimal generator of a strongly continuous cosine family of bounded linear operators, we found sufficient conditions for the existence, uniqueness and continuous dependence of mild and classical solutions on the initial data and bounded perturbations of the impulse operators.

References

1. Akça, H., Boucherif, A., Covachev, V.: Impulsive functional-differential equations with nonlocal conditions. Int. J. Math. Math. Sci. **29**, 251–256 (2002)
2. Akça, H., Covachev, V., Al-Zahrani, E.: On existence of solutions of semilinear impulsive functional differential equations with nonlocal conditions. Oper. Theory Adv. Appl. **153**, 1–11 (2005)
3. Akça, H., Covachev, V., Covacheva, Z.: Existence theorem for a second order impulsive functional-differential equation with a nonlocal condition. J. Nonlinear Convex A. **17**, 1129–1136 (2016)
4. Bochenek, J.: An abstract nonlinear second order differential equation. Ann. Pol. Math. **54**, 155–166 (1991)
5. Byszewski, L., Winiarska, T.: An abstract nonlocal second order evolution problem. Opuscula Math. **32**, 75–82 (2012)
6. Fattorini, H.O.: Ordinary differential equations in linear topological spaces. I. J. Diff. Equat. **5**, 72–105 (1968). II. J. Diff. Equat. **6**, 50–70 (1969)
7. Fattorini, H.O.: Uniformly bounded cosine functions in Hilbert space. Indiana U. Math. J. **20**, 411–415 (1970)
8. Sova, M.: Cosine operator functions. Rozprawy Mat. **49**, 1–47 (1966)
9. Travis, C.C., Webb, G.F.: Cosine family and abstract nonlinear second order differential equations. Acta Math. Acad. Sci. Hungar. **32**, 75–96 (1978)

Representations of Solutions of Hyperbolic Volterra Integro-Differential Equations with Singular Kernels

V. V. Vlasov and N. A. Rautian

Abstract The purpose of the present paper is to study the asymptotic behavior of solutions of integro-differential equations on the basis of spectral analysis of their symbols. To this end, we obtain representations of strong solutions of these equations in the form of a sum of terms corresponding to the real and nonreal parts of the spectrum of the operator functions that are the symbols of these equations. The equations in question are abstract forms of linear partial integro-differential equations arising in the theory of viscoelasticity and in a number of other important applications. These representations are new for the class of integro-differential equations considered in the paper.

1 Introduction

Integro-differential equations with unbounded operator coefficients in a Hilbert space are studied in this work. The equations under consideration are abstract hyperbolic equations perturbed by terms containing Volterra integral operators. The kernels of these Volterra operators are sums of fractional exponential Rabotnov functions. These integro-differential equations can be realized as partial integro-differential equations arising in the theory of viscoelasticity (see [1, 2]) and also as Gurtin-Pipkin integro-differential equations (see [3–5]), which describe heat transfer with a finite rate in media with memory. In addition, equations of this type arise in homogenization problems for multiphase media (see [6, 7]).

V. V. Vlasov (✉) · N. A. Rautian (✉)
Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991,
Russian Federation
e-mail: vikmont@yandex.ru

N. A. Rautian
e-mail: nrautian@mail.ru

Lomonosov Moscow State University, Moscow Center for Fundamental and Applied
Mathematics, Moscow, Russian Federation

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_54

2 Definitions, Notation and Problem Statements

Let H be a separable Hilbert space, and let A be a self-adjoint positive operator $A^* = A$ on H with compact inverse.

$$\frac{d^2u}{dt^2} + A^2u - \int_0^t K(t-s)A^2u(s)\,ds = f(t), \quad t \in \mathbb{R}_+, \tag{1}$$

$$u(+0) = \varphi_0, \tag{2}$$

$$u^{(1)}(+0) = \varphi_1. \tag{3}$$

The scalar function $K(t)$ is representable as

$$K(t) = \sum_{j=1}^{\infty} c_j R_j(t), \tag{4}$$

where $c_j > 0$, $j \in \mathbb{N}$, $R_j(t)$ are fractional exponential Rabotnov functions (see [1, Ch. I]) of the form

$$R_j(t) = t^{\alpha-1} \sum_{n=0}^{\infty} \frac{(-\beta_j)^n t^{n\alpha}}{\Gamma[(n+1)\alpha]}, \quad 0 < \alpha \le 1, \tag{5}$$

$\Gamma(\cdot)$ is the Euler gamma function. We assume that the sequence $\{\beta_j\}$ satisfies the following conditions: $0 < \beta_j < \beta_{j+1}$, $j \in \mathbb{N}$, $\beta_j \to +\infty$, $j \to +\infty$. In addition, we assume that

$$\sum_{j=1}^{\infty} \frac{c_j}{\beta_j} < 1. \tag{6}$$

The Laplace transform of $R_j(t)$ has the form

$$\hat{R}_j(\lambda) = \frac{1}{\lambda^\alpha + \beta_j},$$

(see [2], Ch. I). In this case, λ^α $(0 < \alpha \le 1)$ is understood as the main branch of the multivalued function $f(\lambda) = \lambda^\alpha$, $\lambda \in \mathbb{C}$ with a cut along the negative real half-line: $\lambda^\alpha = |\lambda|^\alpha e^{i\alpha \arg \lambda}$, $-\pi < \arg \lambda < \pi$. Applying the inverse Laplace transform to the main branch of the multivalued function $\hat{R}_j(\lambda)$ we obtain (see [2], Ch. I) the following integral representation of function $R_j(t)$:

$$R_j(t) = \frac{1}{2\pi i} \lim_{R\to+\infty} \int\limits_{\gamma-iR}^{\gamma+iR} \frac{e^{\lambda t} d\lambda}{\lambda^\alpha + \beta_j} = \frac{\sin\pi\alpha}{\pi} \int\limits_0^{+\infty} \frac{e^{-t\tau} d\tau}{\tau^\alpha + 2\beta_j \cos\pi\alpha + \beta_j^2 \tau^{-\alpha}}.$$

Considering the Laplace transform of (1) with homogeneous initial conditions, we arrive at the equation $L(\lambda)\hat{u}(\lambda) = \hat{f}(\lambda)$, where the operator function

$$L(\lambda) = \lambda^2 I + A^2 - \hat{K}(\lambda)A^2, \tag{7}$$

is the symbol of this equation, while $\hat{u}(\lambda)$ and $\hat{f}(\lambda)$ are the Laplace transforms of the vector functions $u(t)$ and $f(t)$, respectively; here, $\hat{K}(\lambda)$ is the Laplace transform of the kernel $K(t)$, which is representable as

$$\hat{K}(\lambda) = \sum_{j=1}^{\infty} \frac{c_j}{\lambda^\alpha + \beta_j}, \quad 0 < \alpha \le 1. \tag{8}$$

In this work, we study the problem of spectrum localization for the operator function $L(\lambda)$ being the symbol of this Eq. (1) and establish the results about representations of the strong solutions of this equation.

In our previous works [8, 9, 11, 13], we studied problem the initial value problem (1)–(3) in detail in the case when the kernel $K(t)$ is representable as a series of decreasing exponentials with positive coefficients, which is equivalent to the case $\alpha = 1$ in representation (4). Our approach was based on the spectral analysis of operator function (7) which also makes it possible to derive a result concerning correct solvability and the representation of the solution of the problem under consideration as a series in exponentials corresponding to spectral points of $L(\lambda)$. We also note that the results of [8, 9, 11] are summarized in Chapter 3 in [10].

3 Formulation of Results

We convert the domain $Dom(A^\beta)$ of the operator A^β, $\beta > 0$, into a Hilbert space H_β by introducing the norm $\|\cdot\|_\beta = \|A^\beta \cdot\|$ on $Dom(A^\beta)$, which is equivalent to the norm of the graph A^β.

Let $W^n_{2,\gamma}(\mathbb{R}_+, A^n)$ denote the Sobolev space of vector-functions on the semiaxis $\mathbb{R}_+ = (0, \infty)$ with values in H equipped with the norm

$$\|u\|_{W^n_{2,\gamma}(\mathbb{R}_+,A^n)} \equiv \left(\int_0^\infty e^{-2\gamma t} \left(\left\|u^{(n)}(t)\right\|^2_H + \left\|A^n u(t)\right\|^2_H \right) dt \right)^{1/2}, \quad \gamma \ge 0.$$

See [15], Ch.1 for more details on the spaces $W^n_{2,\gamma}\left(\mathbb{R}_+, A^2\right)$. For $n = 0$, we put $W^0_{2,\gamma}\left(\mathbb{R}_+, A^0\right) = L_{2,\gamma}\left(\mathbb{R}_+, H\right)$, where $L_{2,\gamma}\left(\mathbb{R}_+, H\right)$ denotes the space of measurable functions with values in H equipped with the norm

$$\|f\|_{L_{2,\gamma}(\mathbb{R}_+,H)} = \left(\int\limits_0^{+\infty} e^{-2\gamma t}\|f(t)\|^2{}_H dt\right)^{1/2}.$$

Definition 1. A vector-function u is said to be a strong solution of problem (1)–(3) if it belongs to $W^2_{2,\gamma}(\mathbb{R}_+, A^2)$ for some $\gamma \geqslant 0$, satisfies (1) almost everywhere on the half-line $\mathbb{R}_+$, and obeys initial condition (2), (3).

3.1 Spectral Analysis

Let a_j denote the eigenvalues of the operator A $(Ae_j = a_j e_j)$ numbered in increasing order: $0 < a_1 < a_2 < \ldots < a_n < \ldots, a_n \to +\infty, (n \to +\infty)$. The corresponding eigenvectors $\left\{e_j\right\}_{j=1}^{\infty}$ form an orthonormal basis in the space H. We consider the restriction of the operator function $L(\lambda)$ to the one-dimensional space spanned by e_n:

$$l_n(\lambda) = (L(\lambda)e_n, e_n) = \lambda^2 + a_n^2\left(1 - \sum_{k=1}^{\infty}\frac{c_k}{\lambda^\alpha + \beta_k}\right). \quad (9)$$

We study the structure of the spectrum of $L(\lambda)$ in the case when condition (6) is hold.

Theorem 1. *Assume that condition* (6) *holds. Then the spectrum of the operator function $L(\lambda)$ is in the open left half-plane.*

Remark 1. When the condition $\sum\limits_{j=1}^{\infty}\frac{c_j}{\beta_j} > 1$ holds, there are infinitely many real eigenvalues of $L(\lambda)$ in the right half-plane. Thus, (6) is a necessary condition for the stability of problem (1)–(3).

If the condition $\sum\limits_{j=1}^{\infty}\dfrac{c_j}{\gamma_j} = 1$ holds, then point $\lambda = 0$ belongs to the spectra of operator function $L(\lambda)$ and it is the eigenvalue of infinite multiplicity.

Theorem 2. *Assume that condition* (6) *holds and $c_j = 0$ for all j above some $N \in \mathbb{N}$. Then the spectra of operator function $L(\lambda)$ is representable as*

$$\sigma(L) := \overline{\left\{\lambda_n^{\pm} \in \mathbb{C}\backslash\mathbb{R}, \lambda_n^- = \overline{\lambda_n^+} | n \in \mathbb{N}\right\}}, \quad (10)$$

where $\lambda_n^{\pm}$ are two nonreal complex conjugate zeros of $L(\lambda)$ that for each sufficiently large $n \in \mathbb{N}$ have the asymptotics

$$\lambda_n^{\pm} = -\sin\left(\frac{\pi\alpha}{2}\right) a_n^{1-\alpha}\frac{Q}{2} \pm i a_n \left(1 - \cos\left(\frac{\pi\alpha}{2}\right) a_n^{-\alpha}\frac{Q}{2}\right) + o\left(a_n^{1-\alpha}\right), \quad n \to +\infty, \tag{11}$$

where $Q = \sum\limits_{j=1}^{N} c_j$.

It is appropriate to make the following important remark.

Remark 2. For $\alpha = 1$ asymptotic formula (11) becomes the previously known asymptotic formula (2.15) from [8] (see also [10]).

Full proofs on the Theorem 2 see in the [12]. Proofs of the close results are contained in [20, 21].

3.2 Representations of the Solutions

We formulate theorems on a representation of the strong solution of problem (1)–(3). Let us introduce the following notations:

$$\mathscr{K}_n(\tau) = \frac{a_n^2\left(\hat{K}_-(-\tau) - \hat{K}_+(-\tau)\right)}{\left(\tau^2 + a_n^2\left(1 - \hat{K}_+(-\tau)\right)\right)\left(\tau^2 + a_n^2\left(1 - \hat{K}_-(-\tau)\right)\right)},$$

$$\hat{K}_{\pm}(-\tau) = \sum_{k=1}^{N} \frac{c_k}{\tau^\alpha e^{\pm i\pi\alpha} + \beta_k}.$$

Theorem 3. *Assume that the assumptions of Theorem 2 hold,* $\alpha \in \left(0, \frac{1}{2}\right)$ *and* $f(t) \equiv 0$. *Then the strong solution of problem* (1)–(3) *is representable in the form*

$$u(t) = u_I(t) + u_R(t), \quad t > 0, \tag{12}$$

where the vector-function $u_I(t)$ *is representable as*

$$u_I(t) = \sum_{n=1}^{\infty} \left(\omega_n(t, \lambda_n^+) + \omega_n(t, \lambda_n^-)\right) e_n, \quad \omega_n(t, \lambda) = \frac{(\varphi_{1n} + \lambda\varphi_{0n})\, e^{\lambda t}}{l_n^{(1)}(\lambda)}, \tag{13}$$

while the vector-function $u_R(t)$ *is representable as*

$$u_R(t) = \sum_{n=1}^{\infty} u_{Rn}(t) e_n, \quad u_{Rn}(t) = \int_0^{\infty} e^{-t\tau} \mathscr{K}_n(\tau)\,(-\tau\varphi_{0n} + \varphi_{1n})d\tau, \tag{14}$$

moreover, series (13) *and* (14) *converge in the norm of the space* H *and* $\lambda_n^{\pm}$ *are nonreal eigenvalues of operator function* $L(\lambda)$ *and* $\varphi_{kn} = (\varphi_k, e_n)$, $n \in \mathbb{N}$, $k = 0, 1$.

Theorems 4 and 5 stated below give estimates of the vector functions $u_I(t)$ and $u_R(t)$. Note that the component $u_I(t)$ corresponds to nonreal eigenvalues $\lambda_n^{\pm}$ and is responsible for the wave nature of the behavior of the solutions. The component $u_R(t)$ is responsible for the behavior of the operator function $L^{-1}(\lambda)$ at the cut along the negative half-line. Thus, representation (12) gives a dichotomy of the solution.

Let P_n denote the orthoprojector onto the subspace being the linear span of the vectors $\{e_j\}_{j=1}^{n}$, and let Q_n denote the orthoprojector onto the subspace orthogonal to the subspace $P_n H$ that is $Q_n = I - P_n$ thus, H is representable as the orthogonal sum

$$H = P_n H \oplus Q_n H.$$

We give results on estimates for the projections of $u_I(t)$ onto $Q_n H$ and $P_n H$.

Theorem 4. *Assume that the assumptions of Theorem 3 hold and the initial data are such that $\varphi_0 \in H_3$ and $\varphi_1 \in H_2$. Then, for any $\varepsilon > 0$ there is a natural number n_0 such that the vector function $u_I(t)$ defined by (13) satisfies the estimates*

$$\left\| Q_{n_0} A^m u_I(t) \right\| \leqslant \theta_1 \left\| Q_{n_0} e^{-kA^{1-\alpha}t} A^m \varphi_0 \right\| + \theta_2 \left\| Q_{n_0} e^{-kA^{1-\alpha}t} A^{m-1} \varphi_1 \right\|, \quad t > 0, \tag{15}$$

$$0 < k = \frac{1}{2} \sin\left(\frac{\pi\alpha}{2}\right) \sum_{j=1}^{N} c_j - \varepsilon, \tag{16}$$

$$\left\| P_{n_0} A^m u_I(t) \right\| \leqslant \theta_3 e^{-\delta t} \left\{ \left\| P_{n_0} A^m \varphi_0 \right\| + \left\| P_{n_0} A^{m-1} \varphi_1 \right\| \right\}, \quad t > 0, \tag{17}$$

$$\delta = \operatorname{dist}\left(\left\{ \lambda_j^{\pm} \right\}_{j=1}^{n_0}, \{ iy | y \in \mathbb{R} \} \right), \tag{18}$$

where $m = 0, 1, 2$, the positive constants δ, θ_1, θ_2, θ_3 do not depend on the vectors φ_0 and φ_1.

Note that δ is the distance from the subset of nonreal eigenvalues $\left\{ \lambda_j^{\pm} \right\}_{j=1}^{n_0}$ to the imaginary axis. Due to Theorem 1 the spectra of the operator function $L(\lambda)$ is in the left half-plane.

Theorem 5. *Assume that the assumptions of Theorem 4 hold. Then, for any $\varepsilon > 0$ the vector-function $w(t)$ defined by (14) satisfies the estimate*

$$\begin{aligned} \left\| A^m u_R(t) \right\|^2 \leqslant e^{-2\varepsilon t} & \left\{ k_1 \left\| A^{m-\alpha} \varphi_0 \right\|^2 + k_2 \left\| A^{m-1-\alpha} \varphi_1 \right\|^2 \right\} \\ + k_3 & \left\{ \varepsilon^{2(2+\alpha)} \left\| A^{m-2} \varphi_0 \right\|^2 + \varepsilon^{2(1+\alpha)} \left\| A^{m-2} \varphi_1 \right\|^2 \right\}, \quad t > 0, \end{aligned} \tag{19}$$

where $m = 0, 1, 2$, and the positive constants k_1, k_2, k_3 do not depend on the vectors φ_0 and φ_1.

4 Proofs of the Theorems 3–5

4.1 Proof of the Theorem 3

Accordingly conversion formula of the Laplace transform, the strong solution $u\left(t\right)$ of the problem (1)–(3) is representable in the form

$$u\left(t\right)=\text{v. p.}\frac{1}{2\pi i}\int\limits_{\gamma-i\infty}^{\gamma+i\infty}L^{-1}\left(\lambda\right)\left(\lambda\varphi_0+\varphi_1\right)e^{\lambda t}d\lambda$$
$$=\lim_{R\to+\infty}\frac{1}{2\pi i}\int\limits_{\gamma-iR}^{\gamma+iR}L^{-1}\left(\lambda\right)\left(\lambda\varphi_0+\varphi_1\right)e^{\lambda t}d\lambda.$$

Let a_j denote the eigenvalues of the operator A $(Ae_j=a_je_j)$ numbered in ascending order: $0<a_1<a_2<\ldots<a_n<\ldots, a_n\to+\infty$, $(n\to+\infty)$. The corresponding eigenvectors $\left\{e_j\right\}_{j=1}^{\infty}$ form an orthonormal basis in the space H. We consider the projection of the vector function $u\left(t\right)$ to the one-dimensional space spanned by e_n: $e_n:u_n\left(t\right)=\left(u\left(t\right),e_n\right)$. Then $u_n\left(t\right)$ is representable as

$$u_n\left(t\right)=\text{v. p.}\frac{1}{2\pi i}\int\limits_{\gamma-i\infty}^{\gamma+i\infty}l_n^{-1}\left(\lambda\right)\left(\lambda\varphi_{0n}+\varphi_{1n}\right)e^{\lambda t}d\lambda$$
$$=\lim_{R\to+\infty}\frac{1}{2\pi i}\int\limits_{\gamma-iR}^{\gamma+iR}l_n^{-1}\left(\lambda\right)\left(\lambda\varphi_{0n}+\varphi_{1n}\right)e^{\lambda t}d\lambda,$$

where $\varphi_{0n}=(\varphi_0,e_n)$, $\varphi_{1n}=(\varphi_1,e_n)$, $l_n\left(\lambda\right)=\lambda^2+a_n^2\left(1-\hat{K}\left(\lambda\right)\right)$. By Theorem 2 the function $\hat{u}_n\left(\lambda\right)$ has simple poles in the left half-plane at the points $\lambda_n^{\pm}$ with a cut along the negative real half-line. On the complex plane we consider the counterclockwise contour Γ where

$$\Gamma=\Gamma_0\cup\Gamma_1\cup C_R^+\cup R^+\cup R^-\cup C_R^-\cup\Gamma_2,$$

$$\Gamma_0=\{\lambda:\operatorname{Re}\lambda=\gamma,-R\leqslant\operatorname{Im}\lambda\leqslant R\},\quad\Gamma_1=\{\lambda:0\leqslant\operatorname{Re}\lambda\leqslant\gamma,\operatorname{Im}\lambda=R\},$$

$$C_R^+=\{\lambda:\lambda=\mathrm{Re}^{i\varphi},\frac{\pi}{2}\leqslant\varphi\leqslant\pi\},\quad R^+=\{\lambda:\operatorname{Im}\lambda=0,-R\leqslant\operatorname{Re}\lambda\leqslant0\},$$

$$R^-=\{\lambda:\operatorname{Im}\lambda=0,-R\leqslant\operatorname{Re}\lambda\leqslant0\},\quad C_R^-=\{\lambda:\lambda=\mathrm{Re}^{i\varphi},-\pi\leqslant\varphi\leqslant\frac{-3}{2}\pi\},$$

$$\Gamma_2 = \{\lambda : 0 \leqslant \operatorname{Re}\lambda \leqslant \gamma, \operatorname{Im}\lambda = -R\}.$$

Remind (see proof of Lemma 3.1 from [18]) that on the complex plane with a cut along the negative real half-line the following estimate holds:

$$\left|\hat{K}(\lambda)\right| \leqslant \frac{const}{|\lambda|^{\alpha}}. \tag{20}$$

For sufficiently large modules λ on the complex plane with a cut along the negative real half-line, we have

$$\left|\frac{\lambda}{\lambda^2 + a_n^2\left(1 - \hat{K}(\lambda)\right)}\right| \leqslant \frac{|\lambda|}{|\lambda|^2\left|1 + a_n^2\left(\frac{1}{\lambda^2} - \frac{\hat{K}(\lambda)}{\lambda^2}\right)\right|} \leqslant \frac{1}{|\lambda|\left|1 - \left|\frac{a_n^2}{\lambda^2} + \frac{a_n^2\hat{K}(\lambda)}{\lambda^2}\right|\right|}. \tag{21}$$

We note that for sufficiently large modules $\lambda = \mathrm{Re}^{i\varphi}$ and for given a_n the following estimate is valid

$$\left|\frac{a_n^2}{R^2 e^{2i\varphi}} + \frac{a_n^2}{R^{(2+\alpha)}e^{i\varphi(2+\alpha)}}\right| \leqslant \frac{a_n^2}{R^2\left(1 + \frac{1}{R^{\alpha}}\right)} \leqslant \frac{2a_n^2}{R^2}. \tag{22}$$

Hence for sufficiently large radius $R_0 > 2a_n$ for all $R > R_0$ we obtain the inequality

$$\frac{2a_n^2}{R^2} < \frac{1}{2}. \tag{23}$$

It follows from the last inequality and estimates (20)–(23) that

$$\left|l_n^{-1}(\lambda)(\lambda\varphi_{0n} + \varphi_{1n})\right| \leqslant \frac{2}{R}|\varphi_{0n}| + \frac{2}{R^2}|\varphi_{1n}|, \tag{24}$$

for all $R > 2a_n$, $\lambda = \mathrm{Re}^{i\varphi}$. It follows from the estimate (24) the Jordan lemma that for any $t > 0$

$$\int\limits_{C_R^{\pm}} l_n^{-1}(\lambda)(\lambda\varphi_{0n} + \varphi_{1n})e^{\lambda t}d\lambda \to 0, \quad |\lambda| = R \to +\infty.$$

Now let us show that the integrals from function $\hat{u}_n(\lambda)$ over the segments Γ_1 and Γ_2 tend to zero as $R \to +\infty$. By using the estimate (24) we obtain the chain of inequalities

$$\left|\int\limits_{\pm iR}^{\pm iR+\gamma} l_n^{-1}(\lambda)\left(\lambda\varphi_{0n}+\varphi_{1n}\right)e^{\lambda t}d\lambda\right| \leqslant \int\limits_{0}^{\gamma}\left|l_n^{-1}(x\pm iR)\left((x\pm iR)\varphi_{0n}+\varphi_{1n}\right)\right|e^{xt}dx$$
$$\leqslant\left(\frac{2}{R}\left|\varphi_{0n}\right|+\frac{2}{R^2}\left|\varphi_{1n}\right|\right)\frac{e^{\gamma t}-1}{t}. \tag{25}$$

It is easy to see that the right part of (25) tend to zero as $R \to +\infty$.

Let us analyze the behavior $\hat{u}_n(\lambda)$ on the upper and lower coasts of the cut respectively. At first consider the upper coast of the cut from $-R$ until $-\varepsilon$. On the upper coast of the cut we obtain the expression, having the following form for $\lambda = x$:

$$\int\limits_{-R}^{-\varepsilon}\frac{e^{xt}\left(x\varphi_{0n}+\varphi_{1n}\right)}{x^2+a_n^2\left(1-\hat{K}_+(x)\right)}dx=\int\limits_{\varepsilon}^{R}\frac{e^{-\tau t}\left(-\tau\varphi_{0n}+\varphi_{1n}\right)}{\tau^2+a_n^2\left(1-\hat{K}_+(-\tau)\right)}d\tau, \tag{26}$$

where function $\hat{K}_+(-\tau)$ have the form

$$\hat{K}_+(-\tau)=\sum_{k=1}^{N}\frac{c_k}{\tau^\alpha e^{i\pi\alpha}+\beta_k}.$$

Now consider the possibility of limit passage for $\varepsilon \to +0$ $R \to +\infty$. Since to the condition of theorem

$$\sum_{k=1}^{N}\frac{c_k}{\beta_k}<1,$$

then under the integral sign expression in formula (26) do not have singularities in the neighborhood of the origin< that is integral will be not singular for $\tau \to +0$ and is possible (26) to put $\varepsilon = 0$. On account of exponential decreasing of integrand for $\tau \to +\infty$ in integral is possible to pass to the limit for $R \to +\infty$. Hence, on the upper coast of the cut after limit procedure we obtain the expression

$$\int\limits_{0}^{+\infty}\frac{e^{-t\tau}\left(-\tau\varphi_{0n}+\varphi_{1n}\right)}{\tau^2+a_n^2\left(1-\hat{K}_+(-\tau)\right)}d\tau. \tag{27}$$

Absolutely analogously we obtain the expression on the lower coast of the cut

$$-\int\limits_{0}^{+\infty}\frac{e^{-t\tau}\left(-\tau\varphi_{0n}+\varphi_{1n}\right)}{\tau^2+a_n^2\left(1-\hat{K}_-(-\tau)\right)}d\tau, \tag{28}$$

where function $\hat{K}_-(-\tau)$ has the form

$$\hat{K}_{-}(-\tau) = \sum_{k=1}^{N} \frac{c_k}{\tau^{\alpha} e^{-i\pi\alpha} + \beta_k}.$$

Let us realize the limit passage in integrating function $\hat{u}_n(\lambda)$ Γ for $R \to +\infty$. Then integral for contour Γ_0 in limit, for $R \to +\infty$, will give function $u_n(t)$, due to inversion formula for Laplace transform. The integrals of function $\hat{u}_n(\lambda)$ for contours $\Gamma_1, \Gamma_2, C_R^+, C_R^-$, will tends to zero as we proved. In turn the integrals along the cuts at limit for, $R \to +\infty$, will converse to the following integrals (27), (28). Hence, after passing to limit for $R \to +\infty$, we obtain, that function $u_n(t)$ will have the following representation:

$$u_n(t) = u_n^+(t) + u_n^-(t) + u_{Rn}(t), \tag{29}$$

where

$$u_n^+(t) = \operatorname*{res}_{\lambda=\lambda_n^+} \left(\hat{u}_n(\lambda) e^{\lambda t}\right) = \frac{\left(\lambda_n^+ \varphi_{0n} + \varphi_{1n}\right) e^{\lambda_n^+ t}}{2\lambda_n^+ - a_n^2 \hat{K}^{(1)}\left(\lambda_n^+\right)},$$

$$u_n^-(t) = \operatorname*{res}_{\lambda=\lambda_n^-} \left(\hat{u}_n(\lambda) e^{\lambda t}\right) = \frac{\left(\lambda_n^- \varphi_{0n} + \varphi_{1n}\right) e^{\lambda_n^- t}}{2\lambda_n^- - a_n^2 \hat{K}^{(1)}\left(\lambda_n^-\right)},$$

$$u_{Rn}(t) = \int_0^{+\infty} \frac{e^{-t\tau}\left(-\tau\varphi_{0n} + \varphi_{1n}\right)}{\tau^2 + a_n^2\left(1 - \hat{K}_{+}(-\tau)\right)} d\tau - \int_0^{+\infty} \frac{e^{t\tau}\left(-\tau\varphi_{0n} + \varphi_{1n}\right)}{\tau^2 + a_n^2\left(1 - \hat{K}_{-}(-\tau)\right)} d\tau. \tag{30}$$

Note, that multiplier $2\pi i$ before residues are absent, because in formula of the inverse formula of Laplace transform occur the multiplier $\dfrac{1}{2\pi i}$.

So, we obtain on formal level, that solution of initial problem admits the representation in the form of sum of the series

$$u(t) = \sum_{n=1}^{\infty} \left(u_n^+(t) + u_n^-(t) + u_{Rn}(t)\right) e_n. \tag{31}$$

The proof of convergence the series (13), (14) and series (30) are obtained in process of proving the Theorem 4 under obtaining the estimates of the sums of these series.

4.2 Proof of the Theorem 4

Let us analyze the convergence of the series

$$\sum_{n=1}^{\infty} \left(u_n^+(t) + u_n^-(t)\right) e_n. \tag{32}$$

Remind that eigenvalues are complex-conjugate $\bar{\lambda}_n^+ = \lambda_n^-$ because, the coefficients of expansion of Rabotnov function are real-valued. Due to the same reason $\overline{\hat{K}^{(1)}\left(\lambda_n^+\right)} = \hat{K}^{(1)}\left(\lambda_n^-\right)$. Hence, the representations are valid

$$u_n^+(t) = \frac{\left(\lambda_n^+\varphi_{0n} + \varphi_{1n}\right) e^{\lambda_n^+ t}\left(2\lambda_n^- - \hat{K}^{(1)}\left(\lambda_n^-\right)\right)}{\left|2\lambda_n^+ - a_n^2\hat{K}^{(1)}\left(\lambda_n^+\right)\right|^2},$$

$$u_n^-(t) = \frac{\left(\lambda_n^-\varphi_{0n} + \varphi_{1n}\right) e^{\lambda_n^- t}\left(2\lambda_n^+ - \hat{K}^{(1)}\left(\lambda_n^+\right)\right)}{\left|2\lambda_n^+ - a_n^2\hat{K}^{(1)}\left(\lambda_n^+\right)\right|^2}. \qquad (33)$$

and consequently,

$$\begin{aligned} & u_n^+(t) + u_n^-(t) \\ = & \frac{\left(4\left|\lambda_n^+\right|^2 \operatorname{Re}(\exp(\lambda_n^+ t)) + 2\operatorname{Re}\left(\exp(\lambda_n^+ t)\lambda_n^+\hat{K}^{(1)}\left(\lambda_n^+\right)\right)\right)\varphi_{0n}}{\left|2\lambda_n^+ - a_n^2\hat{K}^{(1)}\left(\lambda_n^+\right)\right|^2} \\ & + \frac{\left(4\operatorname{Re}\left(\lambda_n^- e^{\lambda_n^+ t}\right) - 2\operatorname{Re}\left(e^{\lambda_n^+ t}\hat{K}^{(1)}\left(\lambda_n^-\right)\right)\right)\varphi_{1n}}{\left|2\lambda_n^+ - a_n^2\hat{K}^{(1)}\left(\lambda_n^+\right)\right|^2}. \end{aligned} \qquad (34)$$

Remark, that at the proof of lemma 3.1 in the article [12], the following estimate was obtained

$$\left|\hat{K}^{(1)}(\lambda)\right| \leqslant \frac{const}{|\lambda|^{\alpha+1}}, \qquad (35)$$

in the domain $\Omega_{\pi-\delta} = \left\{\lambda : |\arg\lambda| < \pi - \delta,\ 0 < \delta < \frac{\pi}{2}\right\}$. Let us use this estimate in order to estimate denominator in the expression (34). We assume for the further description the following proposition.

Proposition 1. *There are exists such positive constants d_1 d_2, that for all n, begining from certain $n_0 \in \mathbb{N}$, the following inequalities are satisfied*

$$d_1 a_n \leqslant \left|\lambda_n^\pm\right| \leqslant d_2 a_n. \qquad (36)$$

This proposition is direct corollary of the asymptotic formula (11). Really, the left part of the inequality (36) arise from evident inequality $\left|\operatorname{Im}\lambda_n^\pm\right| \leqslant \left|\lambda_n^\pm\right|$, and also to the fact,that for $n > n_0$

$$\left|\operatorname{Im} \lambda_n^{\pm}\right| \geqslant \frac{1}{2} a_n \left(1 - \cos \frac{\alpha \pi}{2} a_n^{-\alpha} \frac{Q}{2}\right).$$

Right part of the inequality (36) arise from the following chain of the inequalities, valid for $n > n_0$:

$$\left|\lambda_n^{\pm}\right| \leqslant \left|\operatorname{Re} \lambda_n^{\pm}\right| + \left|\operatorname{Im} \lambda_n^{\pm}\right| \leqslant 2 \left|\operatorname{Im} \lambda_n^{\pm}\right| \leqslant 2a_n.$$

Hence, from the inequalities (35) and (36) we obtain, that for $n > n_0$

$$\left|\hat{K}^{(1)}\left(\lambda_n^{\pm}\right)\right| \leqslant \frac{C_1}{a_n^{1+\alpha}},$$

with certain positive constant C_1. From this estimate and inequality (36) we derive, that for $n > n_0$

$$\left|2\lambda_n^{\pm} - a_n^2 \hat{K}^{(1)}\left(\lambda_n^{\pm}\right)\right| \geqslant \left|2\lambda_n^{\pm}\right| - \left|a_n^2 \hat{K}^{(1)}\left(\lambda_n^{\pm}\right)\right| \geqslant k_1 a_n - k_2 a_n^{1-\alpha} \geqslant C_2 a_n, \quad (37)$$

with certain positive constants k_1, k_2, C_2.

Let us obtain the upper estimate of the numerator in the expression (34), using the inequality (36) and also asymptotic formula (11). For every $\varepsilon > 0$ exists such number $n_0 \in \mathbb{N}$, that for $n > n_0$ numerators of the first and the second terms admit the following estimates respectively

$$\begin{aligned} &\left|4 \operatorname{Re} e^{\lambda_n^+ t} \left|\lambda_n^+\right|^2 \varphi_{0n} - 2 \operatorname{Re}\left(e^{\lambda_n^+ t} \lambda_n^+ \hat{K}^{(1)}\left(\lambda_n^-\right)\right) \varphi_{0n}\right| \\ &\qquad \leqslant C_3 \left(e^{-k a_n^{1-\alpha} t}\left(a_n^2 + C_4 a_n^{-\alpha}\right)\right) \left|\varphi_{0n}\right| \end{aligned} \quad (38)$$

$$\left|2 \operatorname{Re}\left(e^{\lambda_n^+ t}\left(2\lambda_n^- - \hat{K}^{(1)}\left(\lambda_n^-\right)\right)\right) \varphi_{1n}\right| \leqslant C_5 e^{-k a_n^{1-\alpha} t}\left(a_n - C_6 a_n^{-1-\alpha}\right) \left|\varphi_{1n}\right|, \quad (39)$$

with positive constants C_3, C_4, C_5, C_6, where

$$k = \frac{1}{2} \sin\left(\frac{\pi \alpha}{2}\right) \sum_{j=1}^{N} c_j - \varepsilon.$$

Uniting the inequalities (37)–(39), and throwing decreasing terms, on the base (34), we obtain the following estimate

$$\left|u_n^+(t) + u_n^-(t)\right| \leqslant C_7 e^{-k a_n^{1-\alpha} t} \left|\varphi_{0n}\right| + C_8 e^{-k a_n^{1-\alpha} t} a_n^{-1} \left|\varphi_{1n}\right|, \quad (40)$$

with certain positive constants C_7, C_8, independent from $n > n_0$. Hence, on the base (40) for vector-function

$$u_{In}(t) = \sum_{n=n_0+1}^{\infty} \left(u_n^+(t) + u_n^-(t)\right) e_n$$

we receive the following estimate

$$\|u_{In}(t)\|^2 \leqslant \theta_1 \sum_{n=n_0+1}^{\infty} e^{-2ka_n^{1-\alpha}t} |\varphi_{0n}|^2 + \theta_2 \sum_{n=n_0+1}^{\infty} e^{-2ka_n^{1-\alpha}t} a_n^{-2} |\varphi_{1n}|^2 . \quad (41)$$

with positive constants θ_1, θ_2, independent from n, and also from φ_{0n}, φ_{1n}. The estimate (41) may be rewritten in the form

$$\left\| Q_{n_0} u_I(t) \right\|^2 \leqslant \theta_1 \left\| Q_{n_0} e^{-kA^{1-\alpha}t} \varphi_0 \right\|^2 + \theta_2 \left\| Q_{n_0} e^{-kA^{1-\alpha}t} A^{-1} \varphi_1 \right\|^2 , \quad (42)$$

where Q_{n_0} - orthoprojector on the space, orthogonal to finite dimentional subspace, spanned on the vectors $\left\{e_j\right\}_{j=1}^{n_0}$, and constants θ_1, θ_2 which do not depend of φ_1, φ_2. In turn, from the estimate (42) follow the inequality

$$\left\| Q_{n_0} u_I(t) \right\| \leqslant \beta_1 \left\| Q_{n_0} e^{-kA^{1-\alpha}t} \varphi_0 \right\| + \beta_2 \left\| Q_{n_0} e^{-kA^{1-\alpha}t} A^{-1} \varphi_1 \right\| \quad (43)$$

with constants β_1, β_2, independent from φ_0, φ_1.

The estimate (17) of component of the solution $P_{n_0} u_I(t)$ immediately follows from the conclusion of Theorem 1 about the fact, that spectra of operator-function $L(\lambda)$ is lying in open left half-plane. Hence, there is exists such $\delta > 0$, that finite number of eigenvalues $\{\lambda_j^{\pm}\}_{j=1}^{n_0}$ is separate from imaginary axis by vertical strip $\{\lambda : -\delta < \text{Re} < 0\}$. Asymptotic of nonreal eigenvalues $\lambda_j^{\pm}$, given by formula (11), is obtained for sufficiently large a_n. For the description of localization the first $2n_0$ nonreal eigenvalues $\{\lambda_j^{\pm}\}_{j=1}^{n_0}$ we need methods different from asymptotic methods. Hence we obtained the estimates (15), (17) for $m = 0$. In order to receive the cases $m = 1$, $m = 2$ we substitute vector-functions $Au(t)$ and $Au(t)$ instead of $u(t)$ and after that repeat all the steps of the proof. Theorem 4 is proved.

4.3 Proof of the Theorem 5

Let us pass to the estimate of vector-function $u_R(t)$ Consider coordinate functions $u_{Rn}(t) = (u_R(t), e_n)$:

$$u_{Rn}(t) = \int_0^{+\infty} \frac{e^{-t\tau}(-\tau\varphi_0 + \varphi_1) a_n^2 \left(\hat{K}_-(-\tau) - \hat{K}_+(-\tau)\right)}{\left(\tau^2 + a_n^2\left(1 - \hat{K}_+(-\tau)\right)\right)\left(\tau^2 + a_n^2\left(1 - \hat{K}_-(-\tau)\right)\right)} d\tau . \quad (44)$$

Our purpose is the estimate of function $u_{Rn}(t)$.

Note, that

$$\hat{K}_{+}(-\tau)-\hat{K}_{-}(-\tau)=(-\sin\pi\alpha)\,\tau^{\alpha}\sum_{k=1}^{n}\frac{c_k}{(\tau^{\alpha}\cos\pi\alpha+\beta_k)^2+\tau^{2\alpha}\sin^2\pi\alpha}. \quad (45)$$

As far as $\alpha\in\left(0,\dfrac{1}{2}\right)$, then $\cos(\alpha\pi)\geqslant 0$. Using this fact, let us estimate from below the denominator in the expression (44). From the inequalities

$$\left|\tau^2+a_n^2\left(1-\hat{K}_{\pm}(-\tau)\right)\right|\geqslant\tau^2+a_n^2\left(1-\operatorname{Re}\hat{K}_{\pm}(-\tau)\right),$$

$$\begin{aligned}1-\operatorname{Re}\hat{K}_{\pm}(-\tau)&=1-\sum_{k=1}^{n}\frac{c_k\left(\tau^{\alpha}\cos(\alpha\pi)+\beta_k\right)}{\tau^{2\alpha}+2\beta_k\tau^{\alpha}\cos(\alpha\pi)+\beta_k^2}\\&\geqslant 1-\sum_{k=1}^{n}\frac{c_k\left(\tau^{\alpha}\cos(\alpha\pi)+\beta_k\right)}{\tau^{2\alpha}\cos^2(\alpha\pi)+2\beta_k\tau^{\alpha}\cos(\alpha\pi)+\beta_k^2}\\&=1-\sum_{k=1}^{n}\frac{c_k}{\tau^{\alpha}\cos(\alpha\pi)+\beta_k}\geqslant 1-\sum_{k=1}^{n}\frac{c_k}{\beta_k}=\delta^2>0 \quad (46)\end{aligned}$$

with constant $\delta>0$, we obtain

$$\left|\tau^2+a_n^2\left(1-\hat{K}_{\pm}(-\tau)\right)\right|\geqslant\tau^2+\delta^2a_n^2. \quad (47)$$

Hence on the base (46), (47) we obtain the inequality

$$|u_{Rn}(t)|\leqslant\int_0^{\infty}e^{-\tau t}\frac{\left(\tau\,|\varphi_{0n}|+|\varphi_{1n}|\right)a_n^2\left|\hat{K}_{+}(-\tau)-\hat{K}_{-}(-\tau)\right|}{\left(\tau^2+\delta^2a_n^2\right)^2}d\tau=I_{1n}+I_{2n}. \quad (48)$$

Let us devide the integral in the right part (48) in two integrals: from $\varepsilon>0$ to $+\infty$ and from 0 until ε. Denote first integral I_{1n}, second integral denote by I_{2n}. We pass to the estimate of integral I_{1n}. Note that from representation (45) follow, that for $\tau\in(\varepsilon,+\infty)$ the estimate is correct

$$\left|\hat{K}_{+}(-\tau)-\hat{K}_{-}(-\tau)\right|\leqslant\frac{d_1}{\tau^{\alpha}} \quad (49)$$

with positive constant d_1. In addition, for all $\tau\in(\varepsilon,+\infty)$ the estimate is valid

$$\frac{a_n^2}{\tau^2+\delta^2 a_n^2} \leqslant d_2, \tag{50}$$

where d_2 - positive constant independent from n.

On the base of (47)–(50), we receive

$$I_{1n} \leqslant \int\limits_{\varepsilon}^{+\infty} e^{-\tau t} \frac{(\tau\,|\varphi_{0n}| + |\varphi_{1n}|)\, a_n^2 \left|\hat{K}_+(-\tau) - \hat{K}_-(-\tau)\right|}{\left(\tau^2+\delta^2 a_n^2\right)^2} d\tau$$
$$\leqslant e^{-\varepsilon t} q_1 \left\{ \left(\int\limits_{\varepsilon}^{+\infty} \frac{\tau^{1-\alpha}}{\left(\tau^2+\delta^2 a_n^2\right)} d\tau \right) |\varphi_{0n}| + \left(\int\limits_{\varepsilon}^{+\infty} \frac{d\tau}{\tau^{\alpha}\left(\tau^2+\delta^2 a_n^2\right)} \right) |\varphi_{1n}| \right\}, \tag{51}$$

with positive constant q_1, independent from n.

Realizing change of the variables $\eta = \dfrac{\tau}{a_n}$, we obtain

$$\int\limits_{\varepsilon}^{+\infty} \frac{\tau^{1-\alpha}}{\tau^2+\delta^2 a_n^2} d\tau = \frac{1}{a_n^{\alpha}} \int\limits_{\varepsilon/a_n}^{+\infty} \frac{\eta^{1-\alpha}}{\delta^2+\eta^2} d\eta \leqslant \frac{1}{a_n^{\alpha}} \int\limits_{0}^{+\infty} \frac{\eta^{1-\alpha}}{\delta^2+\eta^2} d\eta,$$

$$\int\limits_{\varepsilon}^{+\infty} \frac{d\tau}{\tau^{\alpha}\left(\tau^2+\delta^2 a_n^2\right)} = \frac{1}{a_n^{1+\alpha}} \int\limits_{\varepsilon/a_n}^{+\infty} \frac{d\eta}{\eta^{\alpha}\left(\delta^2+\eta^2\right)} \leqslant \frac{1}{a_n^{1+\alpha}} \int\limits_{0}^{+\infty} \frac{d\eta}{\eta^{\alpha}\left(\delta^2+\eta^2\right)}. \tag{52}$$

Due to the fact, that $\alpha \in \left(0, \dfrac{1}{2}\right)$ integrals in the right part (52) are convergent. It follows from the estimates (51), (52) that

$$I_{1n} \leqslant e^{-\varepsilon t} \left\{ d_1 \left|a_n^{-\alpha} \varphi_{0n}\right| + d_2 \left|a_n^{-(1+\alpha)} \varphi_{1n}\right| \right\} \tag{53}$$

with positive constants d_1, d_2, independent of n.

Integral I_{2n} may be rewritten in the form

$$I_{2n} = \int_0^{\varepsilon} e^{-\tau t} \frac{\left(\tau\left|a_n^{-2}\varphi_{0n}\right| + \left|a_n^{-2}\varphi_{1n}\right|\right) \left|\hat{K}_+(-\tau) - \hat{K}_-(-\tau)\right|}{\left(\frac{\tau^2}{a_n^2}+1\right)^2} d\tau. \tag{54}$$

Note, that for small $\tau \in (0, \varepsilon)$, according to (45)

$$\left|\hat{K}_+(-\tau) - \hat{K}_-(-\tau)\right| \leqslant p_1 \tau^{\alpha} \tag{55}$$

c with positive constant p_1.

In follows from here and (54), (55) that I_{2n} admits the estimate

$$I_{2n} \leqslant p_2 \int_0^{\varepsilon} e^{-\tau t} \left(\tau^{1+\alpha} \left|a_n^{-2}\varphi_{0n}\right| + \tau^{\alpha} \left|a_n^{-2}\varphi_{1n}\right|\right) d\tau = p_2 J_{2n}. \tag{56}$$

with positive constant p_2, independent of n.

Let us change the variables $\theta = \tau t$, $d\theta = t d\tau$. Then integral in right part (56) will has the form:

$$J_{2n} = \frac{1}{t^{2+\alpha}} \int_0^{\varepsilon t} \theta^{1+\alpha} e^{-\theta} \left|a_n^{-2}\varphi_{0n}\right| d\theta + \frac{1}{t^{1+\alpha}} \int_0^{\varepsilon t} \theta^{\alpha} e^{-\theta} \left|a_n^{-2}\varphi_{1n}\right| d\theta. \tag{57}$$

Applying the second mean theorem to integrals in right part (57), we obtain

$$\begin{aligned} J_{2n} = \frac{(\varepsilon t)^{1+\alpha}}{t^{2+\alpha}} \int_{\xi_1}^{\varepsilon t} e^{-\theta} \left|a_n^{-2}\varphi_{0n}\right| d\theta \\ + \frac{(\varepsilon t)^{\alpha}}{t^{1+\alpha}} \int_{\xi_2}^{\varepsilon t} e^{-\theta} \left|a_n^{-2}\varphi_{1n}\right| d\theta, \quad \xi_1, \xi_2 \in (0, \varepsilon t) \end{aligned} \tag{58}$$

In turn integral in right part of the relation (58) admits the following estimate

$$\begin{aligned} J_{2n} \leqslant \frac{(\varepsilon t)^{2+\alpha} \left|a_n^{-2}\varphi_{0n}\right|}{t^{2+\alpha}} + \frac{(\varepsilon t)^{1+\alpha} \left|a_n^{-2}\varphi_{1n}\right|}{t^{1+\alpha}} \\ = \varepsilon^{2+\alpha} \left|a_n^{-2}\varphi_{0n}\right| + \varepsilon^{1+\alpha} \left|a_n^{-2}\varphi_{1n}\right|. \end{aligned} \tag{59}$$

Hence from relations (56) and (59) we receive the estimate

$$I_{2n} \leqslant const \left(\varepsilon^{2+\alpha} \left|a_n^{-2}\varphi_{0n}\right| + \varepsilon^{1+\alpha} \left|a_n^{-2}\varphi_{1n}\right|\right). \tag{60}$$

Uniting the estimates (48), (53) and(60), we obtain as the final result

$$\begin{aligned} |u_{Rn}(t)| \leqslant e^{-\varepsilon t} \left\{d_1 \left|a_n^{-\alpha}\varphi_{0n}\right| + d_2 \left|a_n^{-1-\alpha}\varphi_{1n}\right|\right\} \\ + d_3 \left\{\varepsilon^{2+\alpha} \left|a_n^{-2}\varphi_{0n}\right| + \varepsilon^{1+\alpha} \left|a_n^{-2}\varphi_{1n}\right|\right\}. \end{aligned} \tag{61}$$

From the estimate (61) follows the estimate of vector function $u_R(t)$:

$$\begin{aligned} \|u_R(t)\|^2 \leqslant e^{2\varepsilon t} \left\{k_1 \sum_{n=1}^{\infty} \left|a_n^{-\alpha}\varphi_{0n}\right|^2 + k_2 \sum_{n=1}^{\infty} \left|a_n^{-1-\alpha}\varphi_{1n}\right|^2\right\} \\ + k_3 \left\{\varepsilon^{2(2+\alpha)} \sum_{n=1}^{\infty} \left|a_n^{-2}\varphi_{0n}\right|^2 + \varepsilon^{2(1+\alpha)} \sum_{n=1}^{\infty} \left|a_n^{-2}\varphi_{1n}\right|^2\right\}. \end{aligned} \tag{62}$$

In turn estimate (62) may be rewritten in the form

$$\|u_R(t)\|^2 \leqslant e^{-2\varepsilon t}\left\{k_1\left\|A^{-\alpha}\varphi_0\right\|^2 + k_2\left\|A^{-1-\alpha}\varphi_1\right\|^2\right\} + \\ + k_3\left\{\varepsilon^{2(2+\alpha)}\left\|A^{-2}\varphi_0\right\|^2 + \varepsilon^{2(1+\alpha)}\left\|A^{-2}\varphi_1\right\|^2\right\}.$$

Hence we obtained the estimate (19) for $m = 0$. In order to receive the cases $m = 1$, $m = 2$ we substitute vector-functions $Au(t)$ and $Au(t)$ instead function $u(t)$ and after that repeat all steps of the proof of the case $m = 0$. Theorem 5 is proved.

Acknowledgements This work was supported by Russian Foundation for Basic Research (grant no. 20-01-00288 A).

References

1. Il'yushin, A.A., Pobedrya, B.E.: Osnovy matematicheskoi teorii termovyazkouprugosti (Foundations of Mathematical Theory of Thermoviscoelasticity). Nauka, Moscow (1970)
2. Rabotnov, Yu.N.: Elementy nasledstvennoi mekhaniki tverdykh tel (Elements of Hereditary Mechanics of Solids). Nauka, Moscow (1977)
3. Lykov, A.V.: Problema teplo- i massoobmena (Heat and Mass Exchange Problem). Nauka i Tekhnika, Minsk (1976)
4. Gurtin, M.E., Pipkin, A.C.: Theory of heat conduction with finite wave speed. Arch. Ration. Mech. Anal. **31**, 113–126 (1968)
5. Eremenko, A., Ivanov, S.: Spectra of the Gurtin–Pipkin type equations. SIAM J. Math. Anal. **43**(5), 2296–2306 (2011). https://doi.org/10.1137/100811908
6. Gavrikov, A.A., Ivanov, S.A., Knyaz'kov, D.Yu., Samarin, V.A., Shamaev, A.S., Vlasov, V.V.: Spectral properties of combined media. J. Math. Sci. **164**(6), 948–963 (2010). https://doi.org/10.1007/s10958-010-9776-5
7. Zhikov, V.V.: On an extension of the method of two-scale convergence and its applications. Sb. Math. **191**(7), 973–1014 (2000). https://doi.org/10.1070/SM2000V191N07ABEH000491
8. Rautian, N.A., Vlasov, V.V.: Well-defined solvability and spectral analysis of abstract hyperbolic. J. Math. Sci. **179**(3), 390–415 (2011). https://doi.org/10.1007/s10958-011-0600-7
9. Rautian, N.A., Vlasov, V.V.: Properties of solutions of integro-differential equations arising in heat and mass transfer theory. Trans. Mosc. Math. Soc., 185–204 (2014)https://doi.org/10.1090/S0077-1554-2014-00231-4
10. Rautian, N.A., Vlasov, V.V.: Spektral'nyi analiz funktsional'no-differentsialnykh uravnenii (Spectral Analysis of Functional-Differential Equations). MAKS Press, Moscow (2016)
11. Rautian, N.A., Vlasov, V.V.: Well-posedness and spectral analysis of integrodifferential equations arising in viscoelasticity theory. J. Math. Sci. **233**(4), 555–577 (2018). https://doi.org/10.1007/s10958-018-3943-5
12. Rautian, N.A., Vlasov, V.V.: Well-posedness and spectral analysis of integrodifferential equations arising in viscoelasticity theory. Study of operator models arising in viscoelasticity. Sovrem. Mat. Fundam. Napravl. **64**(1), 60–73 (2018)
13. Vlasov, V.V., Wu, J.: Solvability and spectral analysis of abstract hyperbolic equations with delay. Funct. Diff. Equat. **16**(4), 751–768 (2009)
14. Pandolfi, L.: The controllability of the Gurtin–Pipkin equations: a cosine operator approach. Appl. Math. Optim. **52**, 143–165 (2005)
15. Lions, J.L., Magenes, E.: Nonhomogeneous Boundary-Value Problems and Applications. Springer, Berlin, Heidelberg, New York (1972)

16. Desch, W., Miller, R.: Exponential stabilization of Volterra Integrodifferential equations in Hilbert space. J. Diff. Equat. **70**, 366–389 (1987)
17. Prüss J.: Evolutionary Integral Equations and Applications. Monographs in Mathematics, vol. 87. Birkhauser Verlag, Basel, Baston, Berlin (1993)
18. Wu, J.: Semigroup and integral form of class of partial differential equations with infinite delay. Diff. Integr. Equat. **4**(6), 1325–1351 (1991)
19. Amendola, G., Fabrizio, M., Golden, J.M.: Thermodynamics of Materials with Memory. Theory and Applications. Springer, New-York, Dordrecht, Heidelberg, London (2012)
20. Vlasov, V.V., Rautian, N.A.: Well-posedness and spectral analysis of Volterra integro-differential equations with singular kernels. Dokl. Math. **98**(2), 502–505 (2018)
21. Vlasov, V.V., Rautian, N.A.: Correct solvability and representation of solutions of Volterra integrodifferential equations with fractional exponential kernels. Dokl. Math. **100**(2), 467–471 (2019). https://doi.org/10.1134/S1064562419050211

Chiral Properties of Discrete Joyce and Hestenes Equations

Volodymyr Sushch

Abstract This paper concerns the question of how chirality is realized for discrete counterparts of the Dirac-Kähler equation in the Hestenes and Joyce forms. It is shown that left and right chiral states for these discrete equations can be described with the aid of some projectors on a space of discrete forms. The proposed discrete model admits a chiral symmetry. We construct discrete analogues of spin operators, describe spin eigenstates for a discrete Joyce equation, and also discuss chirality (A preprint version of the article is available as ArXiv preprint: http://arxiv.org/pdf/1912.01296).

Keywords Dirac-Kähler equation · Hestenes equation · Joyce equation · Chirality · Clifford product · Spin eigenstates

1 Introduction

We present some recent results in the discretisation of the Dirac equation in the geometric algebra of spacetime by using the Dirac-Kähler approach. In this approach, a discretisation scheme is geometric in nature and rests upon the use of the differential forms calculus. The general topic of this paper is the description of some discrete constructions in which the chiral properties of the Dirac theory are captured. In the context of the geometric discretisation, it is natural to introduce a Clifford product acting on the space of discrete inhomogeneous forms as was discussed in [20]. This work is a direct continuation of that described in my previous papers [15–20]. In [16], on the issue of chirality, special attention to a discrete Hodge star operator has been paid. A central role of the Hodge star to deal with chiral symmetry in the lattice formulation was already pointed out by Rabin [14]. There are several approaches to study of discrete versions of the Dirac-Kähler equation based on the use of a discrete Clifford calculus framework on lattices. For a review of discrete Clifford analysis, we refer the reader to [5–7, 13, 21].

V. Sushch (✉)
Koszalin University of Technology, Sniadeckich 2, 75-453 Koszalin, Poland
e-mail: volodymyr.sushch@tu.koszalin.pl

S. Pinelas et al. (eds.), *Differential and Difference Equations with Applications*,
Springer Proceedings in Mathematics & Statistics 333,
https://doi.org/10.1007/978-3-030-56323-3_55

We first briefly review some notations and basic facts on the Dirac-Kähler equation [12, 14] and the Dirac equation in the spacetime algebra [8, 9]. Let $M=\mathbb{R}^{1,3}$ be Minkowski space. Denote by $\Lambda^r(M)$ the vector space of smooth complex-valued differential r-forms, $r=0,1,2,3,4$. Let $d:\Lambda^r(M)\rightarrow\Lambda^{r+1}(M)$ be the exterior differential and let $\delta:\Lambda^r(M)\rightarrow\Lambda^{r-1}(M)$ be the formal adjoint of d with respect to the natural inner product in $\Lambda^r(M)$. We have

$$\delta=*d*,$$

where $*$ is the Hodge star operator $*:\Lambda^r(M)\rightarrow\Lambda^{4-r}(M)$ with respect to the Lorentz metric. Denote by $\Lambda(M)$ the set of all differential forms on M. We have

$$\Lambda(M)=\Lambda^0(M)\oplus\Lambda^1(M)\oplus\Lambda^2(M)\oplus\Lambda^3(M)\oplus\Lambda^4(M).$$

Let $\Omega\in\Lambda(M)$ be an inhomogeneous differential form, i.e., $\Omega=\sum_{r=0}^{4}\overset{r}{\omega}$, where $\overset{r}{\omega}\in\Lambda^r(M)$. The Dirac-Kähler equation for a free electron is given by

$$i(d+\delta)\Omega=m\Omega, \tag{1}$$

where i is the usual complex unit and m is a mass parameter.

Let $\{\gamma_0,\gamma_1,\gamma_2,\gamma_3\}$ be a vector basis of the Clifford algebra $C\ell(1,3)$, namely $\gamma_\mu\gamma_\nu+\gamma_\nu\gamma_\mu=2g_{\mu\nu}$, where $g_{\mu\nu}=\mathrm{diag}(1,-1,-1,-1)$ and $\mu,\nu=0,1,2,3$. Hestenes [9] calls this algebra the spacetime algebra. It is known that the vectors γ_μ can be represented by the 4×4 Dirac gamma matrices [1, 9]. Through the identification of the basic covectors dx^μ and the matrices γ_μ which arises from representation theory, one connects the differential forms under the Clifford product to the algebra of gamma matrices. In other words, the graded algebra $\Lambda(M)$ endowed with the Clifford multiplication is an example of a Clifford algebra. It is true that Eq. (1) is equivalent to the four usual Dirac equations (traditional column-spinor equations). Let $\Lambda_{\mathbb{R}}(M)$ denote the set of real-valued differential forms and let $\Lambda^{ev}(M)=\Lambda^0(M)\oplus\Lambda^2(M)\oplus\Lambda^4(M)$. The Dirac equation in the Hestenes form [8, 9] can be written in terms of inhomogeneous forms as

$$-(d+\delta)\Omega^{ev}\gamma_1\gamma_2=m\Omega^{ev}\gamma_0, \quad \Omega^{ev}\in\Lambda^{ev}_{\mathbb{R}}(M). \tag{2}$$

We consider also the generalized bivector Dirac equation [10] in the form

$$i(d+\delta)\Omega^{ev}=m\Omega^{ev}\gamma_0, \quad \Omega^{ev}\in\Lambda^{ev}(M). \tag{3}$$

Following Baylis [2] we call Eq. (3) the Joyce equation. This equation is equivalent to two copies of the usual Dirac equation. For a deeper discussion of equivalence of Dirac formulations we refer the reader to [11].

The goal of this work is to establish the chirality of discrete versions of the Dirac equation in the Hestenes and Joyce forms. We show that defined some projectors on the space of discrete forms one can decompose solutions of Eqs. (1)–(3) into its left-handed and right-handed parts. Two types of such projectors are introduced and we prove that a discrete Dirac-Kähler operator flips the chirality for both of them. We also construct spin $\pm\frac{1}{2}$ eigenstates for a discrete counterpart of the plane wave solution to a discrete Joyce equation and discuss chirality for such fields.

2 Discrete Dirac-Kähler, Hestenes and Joyce Equations

In this section, we recall some discrete constructions concerning the Dirac-Kähler equation and a discrete Clifford calculus. A discretisation scheme is based on the language of differential forms and is described in [16]. The approach was originated by Dezin [4]. For the convenience of the reader we briefly repeat the relevant material from [16] without proofs, thus making our presentation self-contained. All details one can find in [15, 16].

Let $K(4) = K \otimes K \otimes K \otimes K$ be a cochain complex with complex coefficients, where K is the 1-dimensional complex generated by 0- and 1-dimensional basis elements x^{k_μ} and e^{k_μ}, $k_\mu \in \mathbb{Z}$, respectively. Then an arbitrary r-dimensional basis element of $K(4)$ can be written as $s^k_{(r)} = s^{k_0} \otimes s^{k_1} \otimes s^{k_2} \otimes s^{k_3}$, where s^{k_μ} is either x^{k_μ} or e^{k_μ}, $\mu = 0, 1, 2, 3$ and $k = (k_0, k_1, k_2, k_3)$ is a multi-index. The symbol (r) contains the whole required information about the number and position $e^{k_\mu} \in K$ in $s^k_{(r)} \in K(4)$. For example, the 1-dimensional basis elements of $K(4)$ can be written as

$$e^k_0 = e^{k_0} \otimes x^{k_1} \otimes x^{k_2} \otimes x^{k_3}, \qquad e^k_1 = x^{k_0} \otimes e^{k_1} \otimes x^{k_2} \otimes x^{k_3},$$
$$e^k_2 = x^{k_0} \otimes x^{k_1} \otimes e^{k_2} \otimes x^{k_3}, \qquad e^k_3 = x^{k_0} \otimes x^{k_1} \otimes x^{k_2} \otimes e^{k_3}.$$

The 2-dimensional basis elements of $K(4)$ have the form

$$e^k_{01} = e^{k_0} \otimes e^{k_1} \otimes x^{k_2} \otimes x^{k_3}, \quad e^k_{02} = e^{k_0} \otimes x^{k_1} \otimes e^{k_2} \otimes x^{k_3}, \quad e^k_{03} = e^{k_0} \otimes x^{k_1} \otimes x^{k_2} \otimes e^{k_3},$$
$$e^k_{12} = x^{k_0} \otimes e^{k_1} \otimes e^{k_2} \otimes x^{k_3}, \quad e^k_{13} = x^{k_0} \otimes e^{k_1} \otimes x^{k_2} \otimes e^{k_3}, \quad e^k_{23} = x^{k_0} \otimes x^{k_1} \otimes e^{k_2} \otimes e^{k_3}.$$

In the same way one can write down the 3-dimensional basic elements e^k_{012}, e^k_{013}, e^k_{023} and e^k_{123}. Finally, denote by

$$x^k = x^{k_0} \otimes x^{k_1} \otimes x^{k_2} \otimes x^{k_3}, \qquad e^k = e^{k_0} \otimes e^{k_1} \otimes e^{k_2} \otimes e^{k_3}$$

the 0- and 4-dimensional basis elements of $K(4)$.

The complex $K(4)$ is a discrete analogue of $\Lambda(M)$ and cochains play a role of differential forms. Let us call them forms or discrete forms to emphasize their relationship with differential forms. Then we have

$$K(4) = K^0(4) \oplus K^1(4) \oplus K^2(4) \oplus K^3(4) \oplus K^4(4),$$

where $K^r(4)$ denotes the set of all discrete r-forms, and any $\overset{r}{\omega} \in K^r(4)$ can be expressed as

$$\overset{0}{\omega} = \sum_k \overset{0}{\omega}_k x^k, \qquad \overset{2}{\omega} = \sum_k \sum_{\mu<\nu} \omega_k^{\mu\nu} e_{\mu\nu}^k, \qquad \overset{4}{\omega} = \sum_k \overset{4}{\omega}_k e^k, \tag{4}$$

$$\overset{1}{\omega} = \sum_k \sum_{\mu=0}^{3} \omega_k^{\mu} e_{\mu}^k, \qquad \overset{3}{\omega} = \sum_k \sum_{\iota<\mu<\nu} \omega_k^{\iota\mu\nu} e_{\iota\mu\nu}^k, \tag{5}$$

where $\overset{0}{\omega}_k$, $\omega_k^{\mu\nu}$, $\overset{4}{\omega}_k$, ω_k^{μ} and $\omega_k^{\iota\mu\nu}$ are complex numbers.

Let $d^c : K^r(4) \to K^{r+1}(4)$ be a discrete analogue of the exterior derivative d and let $\delta^c : K^r(4) \to K^{r-1}(4)$ be a discrete analogue of the codifferential δ. It is clear that $\delta^c = *d^c*$. For more precise definitions of these operators we refer the reader to [16]. In this paper we give only the difference expressions for d^c and δ^c. Let the difference operator Δ_μ be defined by

$$\Delta_\mu \omega_k^{(r)} = \omega_{\tau_\mu k}^{(r)} - \omega_k^{(r)}, \tag{6}$$

where $\omega_k^{(r)} \in \mathbb{C}$ is a component of $\overset{r}{\omega} \in K^r(4)$ and τ_μ is the shift operator which acts as $\tau_\mu k = (k_0, ...k_\mu + 1, ...k_3)$, $\mu = 0, 1, 2, 3$. For forms (4), (5) we have

$$d^c \overset{0}{\omega} = \sum_k \sum_{\mu=0}^{3} (\Delta_\mu \overset{0}{\omega}_k) e_\mu^k, \qquad d^c \overset{1}{\omega} = \sum_k \sum_{\mu<\nu} (\Delta_\mu \omega_k^\nu - \Delta_\nu \omega_k^\mu) e_{\mu\nu}^k, \tag{7}$$

$$\begin{aligned} d^c \overset{2}{\omega} = \sum_k \big[&(\Delta_0 \omega_k^{12} - \Delta_1 \omega_k^{02} + \Delta_2 \omega_k^{01}) e_{012}^k + (\Delta_0 \omega_k^{13} - \Delta_1 \omega_k^{03} + \Delta_3 \omega_k^{01}) e_{013}^k \\ &+ (\Delta_0 \omega_k^{23} - \Delta_2 \omega_k^{03} + \Delta_3 \omega_k^{02}) e_{023}^k + (\Delta_1 \omega_k^{23} - \Delta_2 \omega_k^{13} + \Delta_3 \omega_k^{12}) e_{123}^k \big], \end{aligned} \tag{8}$$

$$d^c \overset{3}{\omega} = \sum_k (\Delta_0 \omega_k^{123} - \Delta_1 \omega_k^{023} + \Delta_2 \omega_k^{013} - \Delta_3 \omega_k^{012}) e^k, \qquad d^c \overset{4}{\omega} = 0, \tag{9}$$

$$\delta^c \overset{0}{\omega} = 0, \qquad \delta^c \overset{1}{\omega} = \sum_k (\Delta_0 \omega_k^0 - \Delta_1 \omega_k^1 - \Delta_2 \omega_k^2 - \Delta_3 \omega_k^3) x^k, \tag{10}$$

$$\begin{aligned} \delta^c \overset{2}{\omega} = \sum_k \big[&(\Delta_1 \omega_k^{01} + \Delta_2 \omega_k^{02} + \Delta_3 \omega_k^{03}) e_0^k + (\Delta_0 \omega_k^{01} + \Delta_2 \omega_k^{12} + \Delta_3 \omega_k^{13}) e_1^k \\ &+ (\Delta_0 \omega_k^{02} - \Delta_1 \omega_k^{12} + \Delta_3 \omega_k^{23}) e_2^k + (\Delta_0 \omega_k^{03} - \Delta_1 \omega_k^{13} - \Delta_2 \omega_k^{23}) e_3^k \big], \end{aligned} \tag{11}$$

$$\delta^c \overset{3}{\omega} = \sum_k \big[(-\Delta_2\omega_k^{012} - \Delta_3\omega_k^{013})e_{01}^k + (\Delta_1\omega_k^{012} - \Delta_3\omega_k^{023})e_{02}^k \\ +(\Delta_1\omega_k^{013} + \Delta_2\omega_k^{023})e_{03}^k + (\Delta_0\omega_k^{012} - \Delta_3\omega_k^{123})e_{12}^k \\ +(\Delta_0\omega_k^{013} + \Delta_2\omega_k^{123})e_{13}^k + (\Delta_0\omega_k^{023} - \Delta_1\omega_k^{123})e_{23}^k\big], \tag{12}$$

$$\delta^c \overset{4}{\omega} = \sum_k \big[(\Delta_3\overset{4}{\omega}_k)e_{012}^k - (\Delta_2\overset{4}{\omega}_k)e_{013}^k + (\Delta_1\overset{4}{\omega}_k)e_{023}^k + (\Delta_0\overset{4}{\omega}_k)e_{123}^k\big]. \tag{13}$$

Let $\Omega \in K(4)$ be a discrete inhomogeneous form, that is

$$\Omega = \sum_{r=0}^{4} \overset{r}{\omega}, \tag{14}$$

where $\overset{r}{\omega} \in K^r(4)$ is given by (4) and (5). A discrete analogue of the Dirac-Kähler equation (1) can be defined as

$$i(d^c + \delta^c)\Omega = m\Omega. \tag{15}$$

We can write this equation more explicitly by separating its homogeneous components as

$$i\delta^c\overset{1}{\omega} = m\overset{0}{\omega}, \quad i(d^c\overset{1}{\omega} + \delta^c\overset{3}{\omega}) = m\overset{2}{\omega}, \quad id^c\overset{3}{\omega} = m\overset{4}{\omega}, \\ i(d^c\overset{0}{\omega} + \delta^c\overset{2}{\omega}) = m\overset{1}{\omega}, \qquad i(d^c\overset{2}{\omega} + \delta^c\overset{4}{\omega}) = m\overset{3}{\omega}. \tag{16}$$

Substituting (7)–(13) into (16) one obtains the set of 16 difference equations [16].

As in [17], we define the Clifford multiplication of the basis elements x^k and e_μ^k, $\mu = 0, 1, 2, 3$, by the following rules:

(a) $x^kx^k = x^k, \quad x^ke_\mu^k = e_\mu^kx^k = e_\mu^k,$

(b) $e_\mu^ke_\nu^k + e_\nu^ke_\mu^k = 2g_{\mu\nu}x^k, \quad g_{\mu\nu} = \mathrm{diag}(1, -1, -1, -1),$

(c) $e_{\mu_1}^k \cdots e_{\mu_s}^k = e_{\mu_1\cdots\mu_s}^k$ for $0 \le \mu_1 < \cdots < \mu_s \le 3,$

supposing the product to be zero in all other cases.

The operation is linearly extended to arbitrary discrete forms.

Consider the following unit forms

$$x = \sum_k x^k, \qquad e = \sum_k e^k, \qquad e_\mu = \sum_k e_\mu^k, \qquad e_{\mu\nu} = \sum_k e_{\mu\nu}^k, \tag{17}$$

where $\mu, \nu = 0, 1, 2, 3$. Note that the unit 0-form x plays a role of the unit element in $K(4)$, i.e., for any r-form $\overset{r}{\omega}$ we have $x\overset{r}{\omega} = \overset{r}{\omega}x = \overset{r}{\omega}$.

Proposition 1. *For the unit forms $x \in K^0(4)$ and $e_\mu \in K^1(4)$ given by (17) the following holds*

$$e_\mu e_\nu + e_\nu e_\mu = 2g_{\mu\nu}x, \qquad \mu, \nu = 0, 1, 2, 3. \tag{18}$$

Proof. By the rule (b), it is obvious. □

Proposition 2. *Let $\Omega \in K(4)$ be an inhomogeneous discrete form. Then we have*

$$(d^c + \delta^c)\Omega = \sum_{\mu=0}^{3} e_\mu \Delta_\mu \Omega, \tag{19}$$

where Δ_μ is the difference operator which acts on each component of Ω by the rule (6).

Proof. See Proposition 1 in [18]. □

Thus the discrete Dirac-Kähler equation (15) can be rewritten in the form

$$i \sum_{\mu=0}^{3} e_\mu \Delta_\mu \Omega = m\Omega.$$

Let $K^{ev}(4) = K^0(4) \oplus K^2(4) \oplus K^4(4)$ and let $\Omega^{ev} \in K^{ev}(4)$ be a real-valued even inhomogeneous form, i.e., $\Omega^{ev} = \overset{0}{\omega} + \overset{2}{\omega} + \overset{4}{\omega}$. A discrete analogue of the Hestenes equation (2) is defined by

$$-(d^c + \delta^c)\Omega^{ev} e_1 e_2 = m\Omega^{ev} e_0, \tag{20}$$

or equivalently,

$$-\sum_{\mu=0}^{3} e_\mu \Delta_\mu \Omega^{ev} e_1 e_2 = m\Omega^{ev} e_0,$$

where e_1, e_2 and e_0 are given by (17). A discrete analogue of the Joyce equation (3) is given by

$$i(d^c + \delta^c)\Omega^{ev} = m\Omega^{ev} e_0, \tag{21}$$

where $\Omega^{ev} \in K^{ev}(4)$ is a complex-valued even inhomogeneous form. Clearly, Eq. (21) can be rewritten in the form

$$i \sum_{\mu=0}^{3} e_\mu \Delta_\mu \Omega^{ev} = m\Omega^{ev} e_0.$$

Applying (7)–(13) Eqs. (20) and (21) can be expressed also in terms of difference equations (see [17, 18]).

3 Chirality and the Discrete Joyce Equation

In the continuum Dirac theory, the fifth gamma matrix γ_5 defined by $\gamma_5 = i\gamma_0\gamma_1\gamma_2\gamma_3$ plays a central role in formulating chiral fermions. It is known that in the language of differential forms the Hodge star operator $*$ has similar properties, up to sign, as γ_5. The difficulties in defining a discrete Hodge star operator to deal with chirality on the lattice were discussed by Rabin in [14]. Several discrete versions of the Hodge star operator have been proposed in [3, 16, 22] in which the chiral properties for Dirac-Kähler fermions in the geometric discretisation are captured. In this section, we use a discrete analogue of γ_5 to describe the chirality of a discrete Dirac field in the Joyce formulation.

Consider the constant 4-form e_5 defined by

$$e_5 = ie_0e_1e_2e_3 = ie, \tag{22}$$

where $e_\mu \in K^1(4)$ and $e \in K^4(4)$ are given by (17). The form e_5 generates the action $e_5 : \overset{r}{\omega} \to e_5\overset{r}{\omega}$, where $\overset{r}{\omega} \in K^r(4)$. Note also that

$$e_5 : K^r(4) \to K^{4-r}(4).$$

It is easy to check that

$$e_5^2 = x \quad \text{and} \quad e_5e_\mu = -e_\mu e_5 \quad \text{for} \quad \mu = 0, 1, 2, 3. \tag{23}$$

Hence the form $e_5 \in K^4(4)$ has exactly the same properties as γ_5.

Proposition 3. *For any inhomogeneous form $\Omega \in K(4)$ we have*

$$e_5(d^c + \delta^c)\Omega = -(d^c + \delta^c)e_5\Omega. \tag{24}$$

Proof. By Proposition 2 and (23), the equality (24) follows. □

Consider the following constant forms

$$P_L = \frac{x - e_5}{2}, \qquad P_R = \frac{x + e_5}{2}. \tag{25}$$

Since

$$P_L^2 = P_LP_L = P_L, \qquad P_R^2 = P_RP_R = P_R,$$

it follows that P_L and P_R are projectors. Let us represent $\Omega \in K(4)$ as

$$\Omega = \Omega_L + \Omega_R, \tag{26}$$

where

$$\Omega_L = P_L\Omega, \qquad \Omega_R = P_R\Omega. \tag{27}$$

It is clear that $e_5\Omega_R = \Omega_R$ and $e_5\Omega_L = -\Omega_L$. Hence we can say that Ω decomposes into its self-dual and anti-self-dual parts with respect to the action e_5. The self-dual and anti-self-dual components of Ω correspond to the chiral right and chiral left parts of a solution of the discrete Dirac-Kähler equation.

Proposition 4. *If Ω is a solution of the massless discrete Dirac-Kähler equation*

$$i(d^c + \delta^c)\Omega = 0, \tag{28}$$

then so are both Ω_R and Ω_L.

Proof. Let $\Omega \in K(4)$ be a solution of Eq. (28). Using (24) and (27) we obtain

$$i(d^c + \delta^c)(\Omega \pm e_5\Omega) = i(d^c + \delta^c)\Omega \mp e_5 i(d^c + \delta^c)\Omega = 0.$$

□

From Proposition 4 it follows immediately that the massless discrete Dirac-Kähler equation is invariant under the transformation

$$\Omega \longrightarrow \Omega \pm e_5\Omega. \tag{29}$$

In other words, the discrete model admits the chiral symmetry (29) of Eq. (28) with respect to the action e_5.

Proposition 5. *If Ω is a solution of the discrete Dirac-Kähler equation (15) then we have*

$$\begin{aligned} i(d^c + \delta^c)\Omega_L &= m\Omega_R, \\ i(d^c + \delta^c)\Omega_R &= m\Omega_L. \end{aligned}$$

Proof. From (24) it follows that

$$(d^c + \delta^c)P_L\Omega = P_R(d^c + \delta^c)\Omega, \quad (d^c + \delta^c)P_R\Omega = P_L(d^c + \delta^c)\Omega \tag{30}$$

for any $\Omega \in K(4)$. Let Ω be a solution of Eq. (15). By (30), we have

$$i(d^c + \delta^c)\Omega_L = i(d^c + \delta^c)P_L\Omega = P_R(m\Omega) = m\Omega_R$$

and

$$i(d^c+\delta^c)\Omega_R = i(d^c+\delta^c)P_R\Omega = P_L(m\Omega) = m\Omega_L.$$

□

Hence, just as in the continuum case, the operator $i(d^c+\delta^c)$ flips the chirality and the massive discrete Dirac-Kähler equation decomposes into two parts.

Let $\Omega^{ev}\in K^{ev}(4)$ be a complex-valued even inhomogeneous form. Then we have

$$\Omega^{ev} = P_L\Omega^{ev} + P_R\Omega^{ev} = \Omega^{ev}_L + \Omega^{ev}_R,$$

where P_L and P_R are given by (25). The discrete Joyce equation splits into two parts in the following way.

Proposition 6. *If Ω^{ev} is a solution of the discrete Joyce equation (21) then we have*

$$i(d^c+\delta^c)\Omega^{ev}_L = m\Omega^{ev}_R e_0, \tag{31}$$

$$i(d^c+\delta^c)\Omega^{ev}_R = m\Omega^{ev}_L e_0. \tag{32}$$

Proof. The proof is the same as that for Proposition 5. □

Thus the chiral properties are captured for our discrete model.

4 Chirality and the Discrete Hestenes Equation

Recall that the Hestenes equation is a form of the Dirac equation in the real algebra $C\ell_{\mathbb{R}}(1,3)$. The discrete Hestenes equation acts in the space of real-valued even form $K^{ev}(4)$. Unfortunately, to discus the chiral properties of this equation the action (22) makes no sense because the form e_5 defined by (22) is complex-valued. To make sense of the chiral action one must substitute for e_5 a real-valued action. Let us denote by $*_5$ the following transformation

$$*_5 : \overset{r}{\omega} \to e\overset{r}{\omega}e_2e_1, \tag{33}$$

where $\overset{r}{\omega}\in K^r(4)$ and e, e_2, e_1 are given by (17). It is true that $*_5 : K^{ev}(4)\to K^{ev}(4)$.

Proposition 7. *For any inhomogeneous form $\Omega\in K(4)$ we have*

$$(*_5)^2\Omega = \Omega, \quad \textit{and} \quad (*_5e_\mu + e_\mu *_5)\Omega = 0 \quad \textit{for} \quad \mu = 0,1,2,3. \tag{34}$$

Proof. By definition, $e = e_0e_1e_2e_3$ and $e^2 = ee = -x$. Then for any $\overset{r}{\omega}\in K^r(4)$ we have

$$(*_5)^2\overset{r}{\omega} = *_5(*_5\overset{r}{\omega}) = e(e\overset{r}{\omega}e_2e_1)e_2e_1 = x\overset{r}{\omega}x = \overset{r}{\omega}.$$

Since $e \in K^4(4)$ anticommutes with $e_\mu \in K^1(4)$ for $\mu = 0, 1, 2, 3$, i.e., $ee_\mu = -e_\mu e$, the second equality of (34) follows immediately. □

Proposition 8. *Let $\Omega \in K(4)$ be an inhomogeneous form. Then the following holds*

$$(*_5(d^c + \delta^c) + (d^c + \delta^c)*_5)\Omega = 0. \tag{35}$$

Proof. By (34), the proof repeats the proof of Proposition 3. □

From (34) and (35) it follows that to deal with chirality in the case of the discrete Hestenes equation one can take $*_5$.

Proposition 9. *The massless discrete Dirac-Kähler equation is invariant under the transformation*

$$\Omega \longrightarrow \Omega \pm *_5\Omega. \tag{36}$$

Proof. By (35), it is obvious. □

It follows that the discrete model admits a chiral symmetry of the type (36).

Let us consider the following operations

$$P_L^* = \frac{1 - *_5}{2}, \qquad P_R^* = \frac{1 + *_5}{2}. \tag{37}$$

It is easy to check that

$$(P_L^*)^2\Omega = P_L^*\Omega, \qquad (P_R^*)^2\Omega = P_R^*\Omega, \qquad P_L^*P_R^*\Omega = P_R^*P_L^*\Omega = 0$$

for any $\Omega \in K(4)$. Hence, the operations P_L^* and P_R^* are projectors. Then $\Omega \in K(4)$ can be represented as (26), where

$$\Omega_L = P_L^*\Omega, \qquad \Omega_R = P_R^*\Omega.$$

Let $\Omega^{ev} \in K^{ev}(4)$ be a real-valued even inhomogeneous form. Then the forms $\Omega_L^{ev} = P_L^*\Omega^{ev}$ and $\Omega_R^{ev} = P_R^*\Omega^{ev}$ are even and we have

$$\Omega^{ev} = \Omega_L^{ev} + \Omega_R^{ev}.$$

It should be noted that Ω_R^{ev} and Ω_L^{ev} are self-dual and anti-self-dual parts of Ω^{ev} with respect to the action $*_5$. They correspond to the chiral right and chiral left parts of a solution of the discrete Hestenes equation. Similarly, as in the case of the Joyce equation, we have the following decomposition of the discrete Hestenes equation.

Proposition 10. *If Ω^{ev} is a solution of the discrete Hestenes equation (20) then we have*

$$-(d^c + \delta^c)\Omega_L^{ev}e_1e_2 = m\Omega_R^{ev}e_0,$$
$$-(d^c + \delta^c)\Omega_R^{ev}e_1e_2 = m\Omega_L^{ev}e_0.$$

Proof. Using (35) and (37) we obtain

$$(d^c+\delta^c)P_L^*\Omega = P_R^*(d^c+\delta^c)\Omega, \quad (d^c+\delta^c)P_R^*\Omega = P_L^*(d^c+\delta^c)\Omega$$

for any $\Omega \in K(4)$. Therefore the proof repeats the proof of Proposition 5. □

Let us consider the parity operation $P : K^r(4) \to K^r(4)$ defined by

$$P\overset{r}{\omega} = e_0\overset{r}{\omega}e_0, \tag{38}$$

where $\overset{r}{\omega} \in K^r(4)$ and $e_0 \in K^1(4)$ is given by (17). It is clear that $P^2\overset{r}{\omega} = \overset{r}{\omega}$. But the second statement of Proposition 7 is not true. The parity operation (38) changes the chirality of discrete forms in the following way.

Proposition 11. *For any form $\Omega \in K(4)$ we have*

$$P(P_L^*\Omega) = P_R^*(P\Omega), \qquad P(P_R^*\Omega) = P_L^*(P\Omega), \tag{39}$$

where P_L^* and P_R^* are given by (37).

Proof. Since e_0 commutes with e_2e_1 and anticommutes with e it follows immediately. □

Decompose an even inhomogeneous form $\Omega^{ev} \in K(4)$ as follows

$$\Omega^{ev} = \Omega^{ev}_+ + \Omega^{ev}_-,$$

where Ω^{ev}_+ commutes with e_0 and Ω^{ev}_- anticommutes with it, i.e.,

$$e_0\Omega^{ev}_\pm = \pm\Omega^{ev}_\pm e_0. \tag{40}$$

Proposition 12. *Let $\Omega^{ev}_{\pm R} = P_R^*\Omega^{ev}_\pm$ and $\Omega^{ev}_{\pm L} = P_L^*\Omega^{ev}_\pm$. Then we have*

$$P\Omega^{ev}_{+R} = \Omega^{ev}_{+L}, \qquad P\Omega^{ev}_{-R} = -\Omega^{ev}_{-L},$$
$$P\Omega^{ev}_{+L} = \Omega^{ev}_{+R}, \qquad P\Omega^{ev}_{-L} = -\Omega^{ev}_{-R}.$$

Proof. By (38)–(40), we obtain

$$P\Omega^{ev}_{+R} = P(P_R^*\Omega^{ev}_+) = P_L^*(P\Omega^{ev}_+) = P_L^*(e_0\Omega^{ev}_+e_0) = P_L^*(\Omega^{ev}_+e_0e_0) = P_L^*\Omega^{ev}_+ = \Omega^{ev}_{+L}.$$

The same proof remains valid for all other cases. □

5 Discrete Plane Wave Solutions and Spin Eigenstates

Discrete versions of the plane wave solutions to discrete Joyce and Hestenes equations are constructed in [19] and [20]. In this section, we study spin properties of these solutions in the case of the discrete Joyce equation and discus how the chirality is realized for spin eigenstates in our discrete model. Recall a discrete version of the general plane wave solution for the Joyce equation (see for details [19]). Let $\psi \in K^0(4)$ and let

$$\psi = \sum_k (ip_0+1)^{k_0}(ip_1+1)^{k_1}(ip_2+1)^{k_2}(ip_3+1)^{k_3}x^k,$$

where i is the usual complex unit, $p_\mu \in \mathbb{R}$ and $k=(k_0,k_1,k_2,k_3)$ is a multi-index. Let A be the even inhomogeneous form given by

$$\begin{aligned} A = & \, a_1((m-p_0)x + p_1e_{01} + p_2e_{02} + p_3e_{03}) \\ & + a_2((m-p_0)e_{12} + p_2e_{01} - p_1e_{02} + p_3e) \\ & + a_3((m-p_0)e_{13} + p_3e_{01} - p_1e_{03} - p_2e) \\ & + a_4((m-p_0)e_{23} + p_3e_{02} - p_2e_{03} + p_1e), \end{aligned}$$

where $p_0 = \pm\sqrt{m^2+p_1^2+p_2^2+p_3^2}$, $a_\mu = \frac{\alpha_\mu}{m-p_0}$ and α_μ is an arbitrary complex number for $\mu = 1,2,3,4$. Here the even unit forms $x \in K^0(4)$, $e \in K^4(4)$ and $e_{\mu\nu} \in K^2(4)$ are given by (17). Then the most general plane wave solution of Eq. (21) is

$$\Omega^{ev} = A\psi. \tag{41}$$

Let consider a particular case of (41), namely $p_2 = p_3 = 0$. This situation corresponds to one in the continuum case in which the plane wave solution is propagating along only one axis, e.g., x_1. In the continuum case, such solutions for a Dirac generalized bivector equation are described in [10]. Then we have

$$\psi = \sum_k (ip_0+1)^{k_0}(ip_1+1)^{k_1}x^k \tag{42}$$

and

$$\begin{aligned} A = & \, a_1((m-p_0)x + p_1e_{01}) + a_2((m-p_0)e_{12} - p_1e_{02}) \\ & + a_3((m-p_0)e_{13} - p_1e_{03}) + a_4((m-p_0)e_{23} + p_1e). \end{aligned} \tag{43}$$

Let us introduce the following constant 2-forms

$$S_1 = i\frac{1}{2}e_{23}, \qquad S_2 = -i\frac{1}{2}e_{13}, \qquad S_3 = i\frac{1}{2}e_{12}. \tag{44}$$

By definition, we have $e_{12}e_{12} = e_{13}e_{13} = e_{23}e_{23} = -x$ and one may easy calculate that $S_1^2 + S_2^2 + S_3^2 = \frac{1}{2}(\frac{1}{2} + 1)x$. Hence similarly to the continuum case the forms (44) can be interpreted as spin operators for our discrete model and spin eigenstates of $\pm\frac{1}{2}$ along the direction of propagation can be described for the solution (41), where ψ and A are given by (42) and (43).

An easy computation shows that the equations $S_2A = \frac{1}{2}A$ and $S_3A = \frac{1}{2}A$, where A is given by (43), have only trivial solutions, i.e., $a_1 = a_2 = a_3 = a_4 = 0$. However, the equation $S_1A = \frac{1}{2}A$ has an non-trivial solution. Indeed, applying the spin operator S_1 to (43) we obtain

$$\begin{aligned} S_1A = i\frac{1}{2}\big(&a_1(m - p_0)e_{23} + a_1p_1e + a_2(m - p_0)e_{13} - a_2p_1e_{03} \\ &- a_3(m - p_0)e_{12} + a_3p_1e_{02} - a_4(m - p_0)x - a_4p_1e_{01}\big). \end{aligned} \tag{45}$$

Combining (45) with (43) we conclude that $S_1A = \frac{1}{2}A$ if and only if $a_1 = -ia_4$ and $a_2 = -ia_3$. It follows that A can be represented as

$$A = a_1A_1 + a_2A_2, \tag{46}$$

where

$$\begin{aligned} A_1 &= (m - p_0)x + p_1e_{01} + i(m - p_0)e_{23} + ip_1e, \\ A_2 &= (m - p_0)e_{12} - p_1e_{02} + i(m - p_0)e_{13} - ip_1e_{03}, \end{aligned} \tag{47}$$

and a_1, a_2 are arbitrary constant. Since ψ is a 0-form we have $S_1\Omega^{ev} = \frac{1}{2}\Omega^{ev}$, where Ω^{ev} is the plane wave solution (41) and A is given by (46). On other words, Ω^{ev} is an eigenstate corresponding to the eigenvalue $\frac{1}{2}$ of the spin operator S_1.

It is clear that if $A\psi$ is a solution of the discrete Joyce equation then $\bar{A}\psi$, where $\bar{A}$ denotes the complex conjugate of A, is also a solution. It can also be seen that $\bar{A} = \bar{a}_1\bar{A}_1 + \bar{a}_2\bar{A}_2$ satisfies the equation $S_1\bar{A} = -\frac{1}{2}\bar{A}$. Hence, similarly as in continuum case [10] the solutions $A\psi$ and $\bar{A}\psi$, where ψ and A are given by (42) and (46), can be interpreted as spin up and spin down solutions correspondingly.

It should be noted that the chirality is captured for the spin solutions described above. Applying the projectors (25) to the forms A_1 and A_2 given by (47) one can calculate

$$\begin{aligned} P_RA_1 &= \frac{1}{2}(m - p_0 + p_1)(x + e_{01} + ie_{23} + ie), \\ P_LA_1 &= \frac{1}{2}(m - p_0 - p_1)(x - e_{01} + ie_{23} - ie), \\ P_RA_2 &= \frac{1}{2}(m - p_0 + p_1)(e_{12} - e_{02} + ie_{13} - ie_{03}), \\ P_LA_2 &= \frac{1}{2}(m - p_0 - p_1)(e_{12} + e_{02} + ie_{13} + ie_{03}). \end{aligned}$$

Thus we have the following two left and two right chiral states

$$\Omega^{ev}_{1L} = P_L A_1 \psi, \quad \Omega^{ev}_{2L} = P_L A_2 \psi, \quad \Omega^{ev}_{1R} = P_R A_1 \psi, \quad \Omega^{ev}_{2R} = P_R A_2 \psi, \tag{48}$$

where ψ is given by (41). Obviously, as has already been described in Sect. 3 the forms (48) satisfy Eqs. (31) and (32).

References

1. Baylis, W.E. (ed.): Clifford (Geometric) Algebra with Applications to Physics, Mathematics, and Engineering. Birkhäuser, Boston (1996)
2. Baylis, W.E.: Comment on ‘Dirac theory in spacetime algebra’. J. Phys. A: Math. Gen. **35**, 4791–4796 (2002)
3. de Beaucé, V., Sen, S., Sexton, J.C.: Chiral Dirac fermions on the lattice using geometric discretisation. Nucl. Phys. B (Proc. Suppl.) **129–130**, 468–470 (2004)
4. Dezin, A.A.: Multidimensional Analysis and Discrete Models. CRC Press, Boca Raton (1995)
5. Faustino, N., Kähler, U., Sommen, F.: Discrete Dirac operators in Clifford analysis. Adv. Appl. Cliff. Alg. **17**(3), 451–467 (2007)
6. Faustino, N.: Solutions for the Klein-Gordon and Dirac equations on the lattice based on Chebyshev polynomials. Complex Anal. Oper. Theory **10**(2), 379–399 (2016)
7. Faustino, N.: A conformal group approach to the Dirac-Kahler system on the lattice. Math. Methods Appl. Sci. **40**(11), 4118–4127 (2017)
8. Hestenes, D.: Real spinor fields. J. Math. Phys. **8**(4), 798–808 (1967)
9. Hestenes, D.: Spacetime Algebra. Gordon and Breach, New York (1966)
10. Joyce, W.P.: Dirac theory in space time algebra: I. The generalized bivector Dirac equation. J. Phys. A: Math. Gen. **34**, 1991–2005 (2001)
11. Joyce, W.P., Martin, J.G.: Equivalence of Dirac formulations. J. Phys. A: Math. Gen. **35**, 4729–4736 (2002)
12. Kähler, E.: Der innere differentialkül. Rendiconti Matematica **21**(3–4), 425–523 (1962)
13. Kanamori, I., Kawamoto, N.: Dirac-Kähler fermion from Clifford product with noncommutative differential form on a lattice. Int. J. Mod. Phys. A **19**(5), 695–736 (2004)
14. Rabin, J.M.: Homology theory of lattice fermion doubling. Nucl. Phys. B **201**(2), 315–332 (1982)
15. Sushch, V.: A discrete model of the Dirac-Kähler equation. Rep. Math. Phys. **73**(1), 109–125 (2014)
16. Sushch, V.: On the chirality of a discrete Dirac-Kähler equation. Rep. Math. Phys. **76**(2), 179–196 (2015)
17. Sushch, V.: Discrete Dirac-Kähler equation and its formulation in algebraic form. Pliska Stud. Math. **26**, 225–238 (2016)
18. Sushch, V.: Discrete Dirac-Kähler and Hestenes equations. In: Pinelas, S., et al. (eds.) Differential and Difference Equations with Applications. ICDDEA 2015. Springer Proceedings in Mathematics and Statistics, vol. 164, pp. 433–442. Springer, Cham (2016)
19. Sushch, V.: Discrete versions of some Dirac type equations and plane wave solutions. In: Pinelas, S., et al. (eds.) Differential and difference equations with applications. ICDDEA 2017. Springer Proceedings in Mathematics and Statistics, vol. 230, pp. 463–475. Springer, Cham (2018)
20. Sushch, V.: A discrete Dirac-Kähler equation using a geometric discretisation scheme. Adv. Appl. Clifford Alg. **28**(72), 1–17 (2018)
21. Vaz, J.: Clifford-like calculus over lattices. Adv. Appl. Clifford Alg. **7**(1), 37–70 (1997)
22. Watterson, S.: The chiral and flavour projection of Dirac-Kähler fermions in the geometric discretization. Int. J. Geom. Methods Mod. Phys. **5**(3), 345–362 (2008)